On Parameter Setting in Identifying the Same Languages Involved in Different Language Data

Ren Wu
Yamaguchi Junior College
1346-2, Oaza Daido, Hofu city, Yamaguchi prefecture, 747-1232, Japan

Hiroshi Matsuno
Graduate School of Science and Engineering, Yamaguchi University
1677-1, Yoshida, Yamaguchi city, Yamaguchi prefecture, 753-8511, Japan
Tel/Fax : +81-83-933-5697
E-mail: matsuno@sci.yamaguchi-u.ac.jp

Abstract

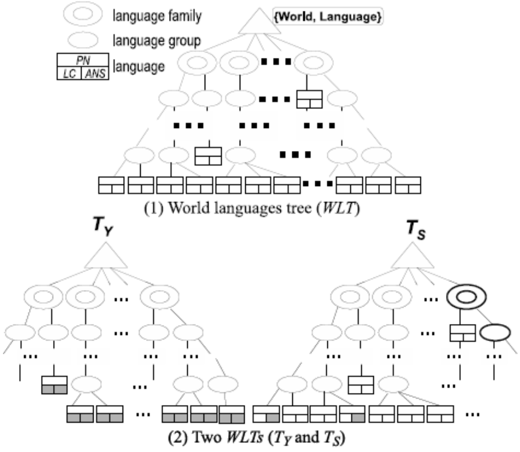
We have introduced several kinds of similarity measure and proposed a method for identifying the same languages involved in two language classification trees. Several unknown parameters are used there and need to be set to constant values. This paper aims to determine all the values of these parameters and get the identification results in order to confirm the usefulness and effectiveness of our proposed similarity measures. As the result, we obtained reasonable good values for the parameters throughout the experiments.

Keywords: world languages tree, language name similarity, language classification similarity

1. Introduction

We have proposed a method based on tree structure and string alignment technique for identifying the same languages involved in two **world languages trees**, denoted by *WLT*s, which are two language classification trees provided by different linguists. We have named these two *WLT*s as T_Y and T_S [1-4].

In our previous work, we have quantified several kinds of similarity measure, such as **language name similarity**, **language classification similarity** and **language general similarity** and so forth. Language name similarity is defined as string similarity between two language names. On the other hand, language classification similarity is defined as a weighted average of three kinds of similarities (family name similarity, parent name similarity and brother name similarity) that are based on language name similarity. Furthermore, language general similarity is defined as a weighted average of language name similarity and language classification similarity. At the same time, we have developed an algorithm for finding out same language pairs involved in T_Y and T_S by applying above defined similarity measures. In this algorithm, several unknown parameters, (e, f, g), (a, b), Δ and threshold ρ, are used and need to be set to constant values firstly.



(1) World languages tree (*WLT*)

(2) Two *WLTs* (T_Y and T_S)

Fig. 1. Two *WLT*s (T_Y and T_S)

This paper aims to determine all the values of these parameters and then get the identification results in order to confirm that our proposed similarity measures and the algorithm are useful and effective. This paper is organized as follows. First, we give the definitions of our proposed similarity measures and the algorithm for finding same language pairs. Then, we give the process for setting the parameters (e,f,g), (a,b), $\varDelta$ and threshold ρ. Finally, we give the experimental results by applying the parameter values that are set.

2. Preliminary

2.1. *Similarity Measure*

Fig. 1(1) shows an image of *WLT* and Fig. 1(2) shows two *WLT*s called T_Y and T_S. A language y included in T_Y is denoted by $y \in V_{leaf}(T_Y)$. Language y and s are assumed to be the same language included in T_Y and T_S respectively.

Fig. 2 shows an example of the same language pair (y,s) involved in T_Y and T_S. The language names of y and s are "ARABIC, SUDANESE CREOLE" and "Arabic, Sudanese Creole" respectively and the same. Here, all alphabetical notations are not case-sensitive.

Finding out such same language pairs like (y,s) automatically is our purpose. In order to solve this problem, measure of language name similarity has been introduced and defined. In the following definition, a term **WORD** is used, which is a string consisting of only alphabets, such as "Arabic". Language name is a list of WORD(s), just like language names "ARABIC, SUDANESE CREOLE" and "Arabic, Sudanese Creole". If two WORDs are individually included in two language names and are most similar, for example "ARABIC" and "Arabic", they are called **WORD pair**.

[**Definition 1**] Let v and w be two WORDs. WORD similarity between v and w, denoted by $sd\_w(v, w)$, is defined by

$$sd\_w(v,w) = \frac{l_A(v,w) - ed(v,w)}{l_A(v,w)}, \qquad (1)$$

where $ed(v, w)$ and $l_A(v, w)$ respectively represent edit distance and length of optimal alignment between v and w, under the condition that the operations of insertions, deletions and substitutions all cost 1 [4]. □

[**Definition 2**] Let $\mathcal{L}_1 = \{ v_1, v_2, \cdots, v_m\}$ and $\mathcal{L}_2 = \{ w_1, w_2, \cdots, w_n\}$ ($m \geqq n$) be language names, (v_i, w'_i) be WORD pair. Note that for any $v_i \in \mathcal{L}_1$, if $\mathcal{L}_2$ does not contain w'_i that corresponds to v_i, then let w'_i = Null. Language name similarity, denoted by $sd\_ln(\mathcal{L}_1,\mathcal{L}_2)$, between $\mathcal{L}_1$ and $\mathcal{L}_2$ is defined by

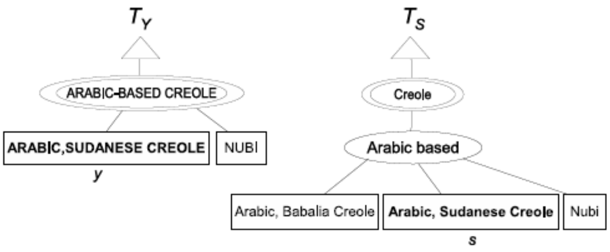


Fig. 2. The same language that needs to be identified

$$sd\_ln(\mathcal{L}_1,\mathcal{L}_2) = \frac{\sum_{i=1}^{m} sd\_w(v_i, w'_i)}{m} \qquad (2)$$

□

[**Definition 3**] Let $y \in V_{leaf}(T_Y)$ and $s \in V_{leaf}(T_S)$ be languages and $\mathcal{L}^{T_Y}_y$ and $\mathcal{L}^{T_S}_s$ be the primary language names of y and s. Let m be the number of alternate names with s, furthermore let all the alternate names be $\mathcal{A}^s_1$, $\mathcal{A}^s_2$, ⋯, $\mathcal{A}^s_m$ if $m > 0$. Then **language name similarity** of y and s, denoted by $sd\_ln_{node}(y, s)$, is defined by

$$sd\_ln_{node}(y,s) = \begin{cases} sd\_ln(\mathcal{L}^{T_Y}_y, \mathcal{L}^{T_S}_s) & (m{=}0) \\ \max\{sd\_ln(\mathcal{L}^{T_Y}_y, \mathcal{L}^{T_S}_s), sd\_ln(\mathcal{L}^{T_Y}_y, \mathcal{A}^s_1), \\ \quad sd\_ln(\mathcal{L}^{T_Y}_y, \mathcal{A}^s_2), \cdots, sd\_ln(\mathcal{L}^{T_Y}_y, \mathcal{A}^s_m)\} & (m{>}0) \end{cases} \qquad (3)$$

□

Based on language name similarity, family name similarity denoted by $sd\_fn(y,s)$, parent name similarity denoted by $sd\_pn(y,s)$ and brother name similarity denoted by $sd\_bn(y,s)$ are defined as follows. These three similarities are designed for language classification similarity, which is intended to be a different similarity measure, from the aspect of language classification.

[**Definition 4**] Let $\mathcal{FN}_y$ and $\mathcal{FN}_s$ be the family names of languages $y \in V_{leaf}(T_Y)$ and $s \in V_{leaf}(T_S)$, respectively. Then family name similarity between y and s, denoted by $sd\_fn(y, s)$, is defined by

$$sd\_fn(y,s) = sd\_ln(\mathcal{FN}_y, \mathcal{FN}_s) \qquad (4)$$

□

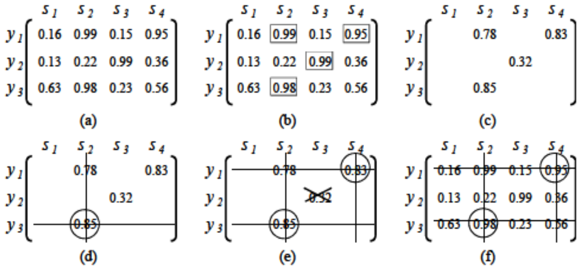


Fig. 3. Process of finding same language pairs

```
Algorithm: FSLV
Input : $V_{leaf}(T_Y)$, $V_{leaf}(T_S)$, $e, f, g, a, b, \Delta, \rho$
Output : Same language pairs $SLP$ for all $y \in V_{leaf}(T_Y)$
Procedure :
1°  Set $SLP \leftarrow \phi$, $P \leftarrow \phi$, $\gamma \leftarrow 1$. For all the pairs of $y \in V_{leaf}(T_Y)$ and $s \in V_{leaf}(T_S)$, do $\psi_{y,s} = sd\_ln_{node}(y, s)$ and generate language name similarity matrix $\Psi = (\psi_{y,s})$.
2°  If $\gamma \leq 0$ goto 3°; otherwise do $P \leftarrow P \cup \{(y, s) \mid \gamma \geq \psi_{y,s} \geq \gamma - \Delta, \psi_{y,s}$: elements of $\Psi\}$ and do the following operations (i) ~ (v):
    (i)   For all $(y, s) \in P$ do $\omega_{y,s} = sd\_gen(y, s)$ by using the coefficients $e, f, g, a, b$.
    (ii)  Do $\omega_{max} = \max\{\omega_{y,s} \mid (y, s) \in P\}$. If $\omega_{max} < \rho$, set $\gamma \leftarrow \gamma - \Delta$ and goto 2°.
    (iii) Select one pair $(y', s') \in P$ with $\omega_{y',s'} = \omega_{max}$ and do $SLP \leftarrow SLP \cup (y', s')$.
    (iv)  Do $P \leftarrow P - \{(y', s')\}$ and delete $y'$'s row and $s'$'s column from $\Psi$.
    (v)   If $P \neq \phi$ goto (ii); otherwise do $\gamma \leftarrow \gamma - \Delta$ and goto 2°.
3°  Output $SLP$ and stop.
```

Fig. 4. Algorithm: FSLV

[**Definition 5**] Let $\mathcal{PN}_y$ and $\mathcal{PN}_s$ be the parent's names of languages $y \in V_{leaf}(T_Y)$ and $s \in V_{leaf}(T_S)$, respectively. Then parent name similarity between y and s, denoted by $sd\_pn(y,s)$, is defined by

$$sd\_pn(y, s) = sd\_ln(\mathcal{PN}_y, \mathcal{PN}_s) \qquad (5)$$

□

Languages "NUBI" and "Nubi" shown in Fig. 2 are brothers of "ARABIC, SUDANESE CREOLE" and "Arabic, Sudanese Creole", respectively. A brother combination from two sets of brothers, such as ("NUBI", "Nubi"), is called **brother language pair**.

[**Definition 6**] Let x_1 (x_2) be a language included in T_Y or T_S (T_S or T_Y), $BL_{x_1}=\{bl^{x_1}_1, bl^{x_1}_2, \cdots\}$ and $BL_{x_2}=\{bl^{x_2}_1, bl^{x_2}_2, \cdots\}$ ($|BL_{x_1}| \geqq |BL_{x_2}| > 0$) be respectively the sets of brother languages of x_1 and x_2, and further $BP(x_1,x_2)$ be the set of brother language pair. Then brother name similarity between x_1 and x_2, denoted by $sd\_bn(x_1, x_2)$, is defined by

		Fact	
		Are two languages the same?	
		Yes	No
Result	Are two languages outputted as the same language pair? Yes	① TP (True Positive)	② FP (False Positive)
	No	③ FN (False Negative)	④ TN (True Negative)

Fig. 5. True-false judgment for outputted same language pairs

$$sd\_bn(x_1, x_2) = \begin{cases} sd\_ln_{node}(x_1, x_2) & (m=n=0) \\ 0 & (m>0, n=0) \\ \dfrac{2\times\sum_{(\mu,\nu)\in BP(x_1,x_2)} sd\_ln_{node}(\mu,\nu)}{m+n} & (m \geq n > 0) \end{cases} \qquad (6)$$

□

Then language classification similarity, and further language general similarity that integrates language name similarity and language classification similarity, are defined as follows.

[**Definition 7**] language classification similarity between $y \in V_{leaf}(T_Y)$ and $s \in V_{leaf}(T_S)$, denoted by $sd\_lc(y, s)$, is defined by

$$sd\_lc(y, s) = e \times sd\_fn(y, s) + f \times sd\_pn(y, s) + g \times sd\_bn(y, s), \qquad (7)$$

where e, f and g satisfy $1 \geqq e \geqq 0$, $1 \geqq f \geqq 0$, $1 \geqq g \geqq 0$ and $e + f + g = 1$. □

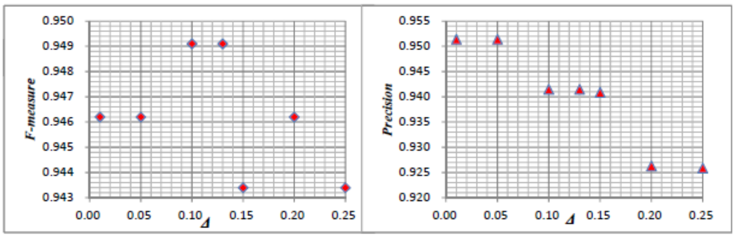
[**Definition 8**] language general similarity between $y \in V_{leaf}(T_Y)$ and $s \in V_{leaf}(T_S)$, denoted by $sd\_gen(y, s)$, is defined by

$$sd\_gen(y, s) = a \times sd\_ln(y, s) + b \times sd\_lc(y, s), \qquad (8)$$

where a and b satisfy $1 \geqq a \geqq 0$, $1 \geqq b \geqq 0$ and $a + b = 1$. □

2.2 *Finding same language pairs*

Same language pairs such as (y,s) shown in Fig. 2 can be found out by applying language name similarity and language general similarity as follows. Two dummy *WLTs* with 3 languages in T_Y and 4 in T_Y are used to describe how to find them.



(e=0.25, f=0.25, g=0.5, a=0.5, b=0.5, ρ=0.5)

(1) $F-measure$ (2) $Precision$

Fig. 6. Ups and downs of the evaluation values

Table 1. Experimental results

TN (total number of languages)	number of languages found out	*TP*	*FP*	*TP*/*TN*	*TP*/ (*TP*+*FP*)
2,869	2,687	2,648	39	**92%**	98%

e=0.25, *f*=0.25, *g*=0.5, *Δ*=0.13, *a*=0.5, *b*=0.5, *ρ* =0.54

(1) First, generate a matrix Ψ with y_i, s_j = $sd\_ln_{node}(y_i, s_j)$ as its element for $y_i \in V_{leaf}(T_Y)$ and $s_j \in V_{leaf}(T_S)$ as shown in Fig. 3 (a).

(2) Then select all the combinations of y_i and s_j satisfying $sd\_ln_{node}(y_i,s_j) > \gamma - \varDelta$. The selected combinations are boxed as shown in Fig. 3 (b). Here, the initial value of and $\varDelta$ are set to 1 and 0.1, respectively.

(3) Calculate the values of $sd\_gen(y_i,s_j)$ for the selected combinations as shown in Fig. 3 (c). Values of coefficients (e, f, g) and (a, b) will be determined by experiments.

(4) Select the pair (y_3,s_2) with the highest language general similarity value from all the values of $sd\_gen(y_i, s_j)$. If we set ρ=0.5, this language pair (y_3, s_2) will be passed over and identified as the same language. Then the language general similarity values as shown Fig. 3 (d) are deleted. Similarly, (y_1, s_4) is identified as well as (y_3,s_2) as shown in Fig. 3 (e).

(5) Delete the values related to the identified same language pairs as shown in Fig. 3 (f).

(6) Update γ as $\gamma = \gamma - \varDelta$ and repeat this process from (2) ～ (5) till $\gamma \leqq 0$. The algorithm of this processing is shown in Fig. 4.

3. Parameter Setting

In algorithm FSLV shown in Fig. 4, parameters (e, f, g), (a,b), $\varDelta$ and threshold ρ are used as input arguments, and are unknown. What should be done first is setting their values. In this section, we are going to give a solution for parameter setting.

3.1 *Evaluation method and test data*

Parameters (e, f, g) are used to calculate language classification similarity which is defined as a weighted average of family name similarity, parent name similarity and brother name similarity, and need to be set at first. These three parameters are individually the coefficients of the three similarities. Take (y,s) shown in Fig. 2 as the example, we consider that, if there are same language pairs such as ("NUBI", "Nubi") within the brothers, then it is most probably that y and s may also be the same language. Hence, we lay weight on g that is the coefficient of brother language similarity, and consider that e=0.25, f=0.25, g=0.5 should be appropriate.

On the other hand, parameters (a,b), $\varDelta$ and threshold ρ will be determined throughout experiments. We use test data of 200 languages which were selected out of all 2,869 languages included in T_Y in a random manner. In advance of setting parameters, we investigate the corresponding languages in T_S for all the languages of this test data. However, it does not mean that the corresponding same languages can always be found in T_S for all the languages.

For the test data, we repeat experiments according to algorithm FSLV shown in Fig. 4 by changing the values of parameters (a,b), $\varDelta$ and threshold ρ, and then check the results with the investigated facts to get the values of TP, FP, TN and FN as shown in Fig. 5. Here, $TP + FP + TN + FN = 200$ should be satisfied.

We use evaluation measures *F-measure*, *Recall*, *Precision* [5] to choose the best (a,b), $\varDelta$ and threshold ρ. These measures are defined as follows.

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$F-measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{11}$$

We consider that the values of parameters (a,b), $\varDelta$ and threshold ρ, with the highest values of *F-measure* and *Precision*, should be the best ones.

3.2 *Process of parameter setting*

Here, we determine the values of parameters (a, b), $\varDelta$ and threshold ρ. Firstly, we deal with parameters (a, b). Repeat executing FSLV with $\varDelta$=0.01 and ρ=0.5 by changing the values of parameters (a, b) by steps of 0.1 as follows. (1.0, 0.0), (0.9, 0.1), ⋯ , (0.1, 0.9), (0.0, 1.0). Then do the same thing as well by increasing the values of $\varDelta$ from 0.01 to 0.25 by steps of 0.01 such as $\varDelta$=0.02, 0.03, ⋯, 0.25. Here, threshold ρ is set as ρ=0.5.

After tallying the values of *TP*, *FP*, *TN* and *FN* for every execution of algorithm FSLV, compute the values of *Recall*, *Precision* and *F-measure* and use them to estimate the trend of parameters (a, b). We get to notice that the values of *F-measure* and *Precision* have the best values when a = 0.5 and b = 0.5. So we set the values of parameters (a, b) to (0.5, 0.5).

What need to be set continuously is the value of parameter $\varDelta$. As described above, we repeat executing algorithm FSLV by increasing the values of $\varDelta$ from 0.01 to 0.25 by steps of 0.01. The values of *F-measure* and *Precision* have ups and downs due to the change of $\varDelta$. We show a part of values of *F-measure* and *Precision* in Fig. 6(1) and Fig. 6(2) respectively by setting $\varDelta$ as 0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.13 under the condition of a=0.5, b=0.5, ρ=0.5. Fig. 6(1) indicates that *F-measure* has a peak at $\varDelta$=0.10 and 0.13. At the two points, Fig. 6(2) has the same highest values. We cannot differentiate the effect between the cases when $\varDelta$=0.10 and $\varDelta$=0.13, So we set $\varDelta$=0.10 or $\varDelta$=0.13.

Finally, we set the value of threshold ρ. Under the condition of $\varDelta$=0.10, a=0.5, b=0.5 or $\varDelta$=0.13, a=0.5, b=0.5, we repeat execution of algorithm FSLV by increasing the value of ρ from 0.50 to 0.80 by steps of 0.05, then we found that *F-measure* has a peak at ρ=0.55. So we repeat the execution additionally by setting ρ as 0.51, 0.52, 0.53 or 0.54 individually. *F-measure* shows the same value at ρ=0.54 or ρ=0.55 as well as *Precision*. Under the condition of the same *F-measure* and *Precision*, the threshold with lower value is better. Therefore, we set ρ=0.54.

Up to here, we have gotten the best values of parameters (a, b), $\varDelta$ and threshold ρ under the condition of e=0.25, f=0.25, g=0.5 by using the test data. As the result, $\varDelta$=0.10, a=0.5, b=0.5, ρ=0.54 or $\varDelta$=0.13, a=0.5, b=0.5, ρ=0.54 are obtained.

4. Experimental Results

We have done experiment by applying the obtained parameters to original data T_Y and T_S to confirm the accuracy of parameter setting and the usefulness and effectiveness of our proposed method. By setting the parameters to (i) $\varDelta$=0.10, a=0.5, b=0.5, ρ=0.54 and (ii) $\varDelta$=0.13, a=0.5, b=0.5, ρ=0.54 (both under e=0.25, f=0.25, g=0.5), we have executed algorithm FSLV for T_Y and T_S and gotten the results shown in Table 1. in all the 2,869 languages of T_Y, we have gotten 2,687 same language pairs as the output of algorithm FSLV, in which 2,648 languages (92%) are the true cases.

5. Concluding Remarks

We have given a way of parameter setting for our previously proposed method. Using the parameter values, about 92% languages of T_Y have been identified, and the precision of identification is up to 98%. These results imply that our parameter setting described above is reasonably good. And at the same time, we have confirmed that our proposed method previously is useful and effective.

It should be noticed that the values of parameters (a, b), $\varDelta$ and threshold ρ are set under the condition of e=0.25, f=0.25, g=0.5. As the future work, we need to do further tests to confirm the parameter setting.

Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Challenging Exploratory Research Numbers 23650129.

References

1. R. Wu and H. Matsuno, Identifying Same Languages by Considering Similarities of Language Name and Language Classification, *Proceedings of The 2010 International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2010)*, pp.516-519.
2. R. Wu, H. Inui and H. Matsuno, New Measurement of Similarity of Language Classification, *Proceedings of The 2012 International Technical Conference on Circuits/Systems, Computers and Communications* (ITC-CSCC 2012), in CD-ROM (F-M2-05).
3. Euzenat, J. and Shvaiko, P.: *Ontology Matching* (Springer-Verlag, 2007)
4. G. Navarro, A guided tour to approximate string matching, *ACM Computing Surveys* (CSUR), 33(1)(2001), pp.31-88.
5. V. Vapnik, *The nature of statistical learning theory* (Springer, New York, 1995)