Interlinking Multilingual LOD Resources: A Study on Connecting Chinese, Japanese, and Korean Resources Using the Unihan Database

Saemi Jang, Satria Hutomo, Soon Gill Hong, and Mun Yong Yi

Department of Knowledge Service Engineering, KAIST, Republic of Korea
Sammy1221@kaist.ac.kr, satriahj@kaist.ac.kr, soonhong@kaist.ac.kr, munyi@kaist.ac.kr

Abstract. This study proposes a novel method with which Chinese, Japanese, and Korean (CJK) resources on the Web can be effectively matched and connected. The three countries share Chinese characters even though Japan and Korea have their own language. Utilizing the Unihan database, which covers more than 45,000 characters commonly used by the three countries, we show that the proposed method outperforms the traditional method based on string matching in finding similar characters and words used in these countries. The results represent a first step towards overcoming the multilingual barrier in semantically interlinking Asian LOD resources.

Linked Open Data (LOD) is an international endeavor to interlink structured data on the Web and create the Web of Data on a global level. Linking data can be achieved by understanding the semantic relationships between data and building explicit links for them. Hence, semantically matching and connecting resources in different languages is crucial to successfully building linked open data around the world.

Approximately 60 percent of the world population is Asians. Resolving multilingual issues for the Asian population is one of the important yet challenging tasks as Asian countries mostly use their own writing systems. Those approaches that have been developed for English alphabets and Western language systems cannot be readily adapted to Asian languages systems as their writing systems are based on different assumptions and conventions. Most of the LOD frameworks have focused on Western language resources and most of the open resources in the LOD cloud are connected to the West, significantly hampering the effort to make the LOD cloud truly a global data space.

In this study, we propose a novel method for matching and interlinking Asian LOD resources and then empirically validate the proposed method using Silk Workbench, an application developed in conjunction with the LOD2 EU-FP7 project[1]. China, Japan, and Korea, shortened as CJK, are geographically close and collectively account for the largest population in Asia. The three countries have had mutual interactions for over a thousand years influencing each other's language system. In particular, Japan and Korea have been affected by the Chinese ideographic characters (Han Chinese), which were used by the Han race a long time ago, which still has a strong impact on the Han Chinese characters used in CJK. Our work exploits the fact that these three countries share the origins and semantics of certain characters even though those characters have developed into often differently looking characters over time.

[1] http://lod2.eu

adfa, p. 1, 2011.

Our Proposed Approach.

The Unihan database is a repository for the Unicode Consortium's collective knowledge regarding the CJK Unified Ideographs. The database contains mapping data to allow conversion to and from other coded character sets and additional information about radical-stroke counts and phonetic information [1][2]. The database represents a character as a 16-bit character code and covers more than 45,000 codes. We used reading and semantic information available from the Unihan database. The reading information in Unihan database shows the pronunciations of the same unified Ideographs in China, Japan and Korea. The semantic information in Unihan includes a variety of possible alternative variants beyond the one-to-one matching of Chinese characters.

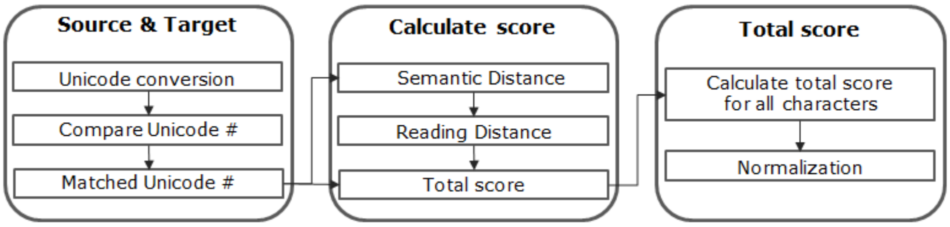

Fig. 1. Process diagram of Han edit distance

To identify matching CJK resources characters using the Unihan database, we propose a new distance measure, called Han Edit Distance (HED). Figure 1 summarizes the overall procedure for the computation of the proposed Han Edit Distance. First, the source and target Chinese words are converted into the Unicode number of each character. Then the two Unicode numbers are compared in a fixed order. It means that the Unicode number of the first character in the source word is compared with the Unicode number of the first character in the target word, the Unicode number of the second character in the source word with the Unicode number of the second character in the target word, and so on. If matching Unicode numbers are found between the characters of those two words, those matching numbers are given a score of 0. If there is not any matching, each Unicode number is converted into the Unicode number of its compatible variant properties. The next step is to check each radical stroke index properties. When the radical part of one character is the same as the other, the two characters belong to the same family. In this case, the number of different strokes is calculated. In the other case, the semantic distance (SD) is given the maximum score of 30 (which is the maximum number of strokes for common Chinese characters) and the reading distance (RD) is calculated by using the reading properties. The total distance is the minimum score between SD and RD. Finally, the Han edit distance is calculated for the two words and then it is normalized. The edit distance calculation algorithm is further detailed below.

Han Edit Distance Calculation Algorithm.

$$\mathrm{HED}\ (0, 0) = 0$$

$$\mathrm{SD}\ (i_1, i_2) = |\ \mathrm{TotalStroke}\ i_1 + \mathrm{TotalStroke}\ i_2\ |$$

$$\mathrm{RD}\ (i\ ,\ j) = [\ \mathrm{M}\ (i, j) + \mathrm{JK}\ (i, j) + \mathrm{JO}\ (i, j) + \mathrm{K}\ (i, j) + \mathrm{H}\ (i, j)\ ] * 6$$

$$\mathrm{HED}\ (s, t) = \min\ [\ \mathrm{SD}\ (s, t)\ ,\ \mathrm{RD}\ (s, t)\ ]$$

Han edit distance is calculated using the distance of both Semantic properties and Reading properties. When two words have the same family root, their semantic distance is calculated. Otherwise, a fixed maximum is assigned to SD, and their reading distance is calculated. Semantic distance represents the difference in the total number of different strokes from the two characters. Although the strokes for each Chinese character are different, the number of Chinese characters that have more than 30 strokes is around 0.23 percent in the Unihan database, so we defined 30 as the maximum number of the total stroke.

Reading distance mainly focuses on how characters are pronounced in each country. When each value of the reading properties is equal, 0 is given as the score; otherwise, 1 is given as the score. Then, the scores from all reading properties are added and multiplied by 6. Multiplying by 6 standardizes RD and SD because their maximum values become very close.

$$\text{Normalized Distance} = 1 - \frac{Distance}{\left(L_i + L_j\right) \times n} \qquad (1)$$

Normalized distance (ND) is defined per Equation 1[3], ranging between 0 and 1 (0≤ND≤1). The sum of the length of elements is multiplied by 30 (n) since the maximum difference between characters is defined as 30. When the Normalized distance is 1, the two words are exactly the same. L_i denotes the length of the source word and L_j denotes the target word. Table 1 shows some examples of Han edit distance between Chinese, Japanese, and Korean words.

Table 1. Examples of Han edit distance between CJK words

	Word	Han	Meaning	Unicode#	SD	RD	HED	ND
1	國家	K	Nation	U+570B U+5BB6	0	30	0	0
	国家	C	Nation	U+56FD U+5BB6				
2	今日	J	Today	U+4ECA U+65E5	0	30	30	0.75
	今天	C	Today	U+4ECA U+5929				
3	読書	J	Reading	U+8AAD U+66F8	0	24	24	0.8
	读书	C	Reading	U+8BFB U+4E66				

Evaluation.

Korea and Japan commonly use about two thousands Chinese characters, and China commonly uses about 2500 characters. Han Chinese words can be composed of just one character or more than one characters. We evaluated the performance of the Han edit distance in two scenarios to reflect this situation. First, similarities at the character level were evaluated utilizing the most commonly used Chinese characters in CJK. 1,937 synonymous pairs were selected as a test data set for this purpose. Second, similarities at the word level were evaluated. 618 pairs of words, each pair of which has the same meaning across the three countries, were selected as a test data set for this purpose. We evaluated our approach against the Levenshtein edit distance, which is most widely used for measuring string similarities. When one character is different, the distance is 1 by the Levenshtein distance while the distance is 30 by the Han edit distance (HED). The two distance measures are compared after normalization.

Table 2. Results from the character-level comparison

Threshold	Chinese : Japanese			Chinese : Korean			Japanese : Korean		
=1	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Levenshtein	0.5975	0.1946	0.2936	0.5919	0.1850	0.2819	0.6065	0.2525	0.3565
HED	0.5982	0.2710	0.3730	0.6477	0.2952	0.4055	0.6012	0.2739	0.3763

Table 3. Result from the word-level comparison

Threshold	Chinese : Japanese			Chinese : Korean			Japanese : Korean		
=1	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Levenshtein	0.9726	0.3430	0.5071	0.9770	0.3414	0.5060	0.9601	0.5427	0.6934
HED	0.9705	0.4767	0.6393	0.9737	0.5958	0.7393	0.9644	0.6538	0.7793

As shown, most scores of the Han edit distance both at the character-level and at the word-level are higher than the scores of the Levensthtein distance. In particular, in all recall and F-measure comparisons, the HED approach shows superior performance consistently at the character-level and at the word-level, without exception. In the comparisons of precision, the results are mixed, but without much noticeable difference between the two approaches. On average, the f-score improvement made by the HED approach is 25% for the character-level comparison and 26% for the word-level comparison.

Concluding Remarks.

Research on LOD mainly focused on Western resources and measured the similarity of the resources at the string level. However, these approaches are not readily applicable to non-Western resources. In this study, we propose a new method to measure similarities among CJK resources, and demonstrate its effectiveness at the character-level and word-level. The results show that the proposed approach is able to identify similar resources of the three countries more effectively than the traditional Levenshtein approach. Our research represents a first step to overcoming the limitations of interlinking multilingual resources in Asia, in particular for CJK. The proposed comparator is planned to be implemented on the next version of Silk Workbench. Future research should involve expanding the approach to include other Asian countries whose characters are covered by the Unihan database such as Singapore, Taiwan, Hong Kong, and Vietnam.

Acknowledgements.

This research was conducted by the International Collaborative Research and Development Program (Creating Knowledge out of Interlinked Data) and funded by the Korean Ministry of Knowledge Economy.

References.

1. The Unicode Standard (http://www.unicode.org/versions/Unicode6.2.0/ch12.pdf).
2. Unihan database document (http://www.unicode.org/charts/unihan.html).
3. Shigeaki Kodama: String Edit Distance for Computing Phonological Similarity between Words, proceedings of International Symposium on Global Multidisciplinary Engineering, (2010).