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

1. Introduction

Background

- 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.
- 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.

Motivation

- 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 thousand years influencing each other's language system.
- Three countries share the origins and semantics of certain characters even though those characters have developed into often differently looking characters over time.

Korea : 만리장성
만리장성(萬里長城)은 흉노족 등의 북방 민족을 막기 위해 중국 진나라에 세워진 거대한 성곽이다.

Japan : ばんりのちょうじょう
ばんりのちょうじょう (万里の長城)は、中華人民共和国にある城壁の遺跡である。

China : 万里长城
长城，是不同时期古代中国为抵御不同时期塞北游牧部落联盟侵袭，修筑规模浩大的军事工程的统称。

2. Methodology

The Unihan database is a repository for the Unicode Consortium's collective knowledge regarding the CJK Unified Ideographs. To identify matching CJK resource characters using the Unihan database, we propose a new distance measure, called Han Edit Distance (HED).

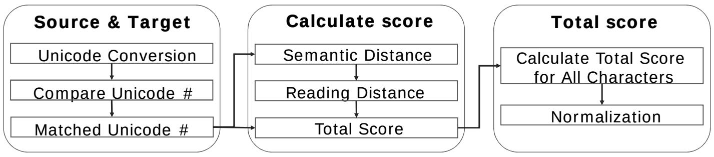
Fig.1. The block construction type of Unihan database

Table.1. The Classification of property in HED

Category	Reading	Semantic
properties	kMandarin kJapaneseOn kJapaneseKun kKorean kHangul	kSemanticVariant kCompatibilityVariant kSimplifiedVariant kRSUnicode kTotalStroke

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.2. Process diagram of Han edit distance



Han edit distance is calculated using the distance of both Semantic properties and Reading properties.

$$\text{HED}(0, 0) = 0$$
$$\text{SD}(i_1, i_2) = |\,\text{TotalStroke } i_1 + \text{TotalStroke } i_2\,|$$
$$\text{RD}(i, j) = [\, M(i, j) + JK(i, j) + JO(i, j) + K(i, j) + H(i, j)\,] * 6$$
$$\text{HED}(s, t) = \min [\, \text{SD}(s, t), \text{RD}(s, t)\,]$$

Fig.3. Han edit distance calculation algorithm

Word	Han	Meaning	Unicode #	SD	RD	HED	ND
萬里長城	Korea	The Great Wall of China	842C,91CC,9577,57CE	0	42	0	1
万里长城	China	The Great Wall of China	4E07, 91CC,957F,57CE				
今日	Japan	Today	4ECA,65E5	30	30	30	0.75
今天	China	Today	4ECA,5929				

Table.2. Example of Han edit distance between CJK words

3. 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.

Scenario 1. Character level
1,937 synonymous pairs, commonly used Han Chinese characters in CJK

Scenario 2. Word level
618 pairs of words, the same meaning across the three countries

Table.3. Result from the character-level comparison

Th=1	Chinese : Japanese			Chinese : Korean			Japanese : Korean		
	Pre	R	F	P	R	F	P	R	F
LD	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.4. Result from the word-level comparison

Th=1	Chinese : Japanese			Chinese : Korean			Japanese : Korean		
	P	R	F	P	R	F	P	R	F
LD	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

Th=Threshold, LD=Levenshtein distance, HED=Han edit distance, P=Precision, R=Recall, F=F-measure

4. Conclusion

- 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.
- 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.
- 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.

5. Acknowledgement

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.