Another Investigation of an Interontologia between Chinese Lexical Systems and Roget's Thesaurus

Sang-Rak Kim, Jae-Gun Yang, and Jae-Hak J. Bae*

School of Computer Engineering & Information Technology, University of Ulsan, Ulsan, Republic of Korea
shem0304@gmail.com, {jgyang,jhjbae}@ulsan.ac.kr

Abstract. The present study presents the lexical category relevancy analysis of the Thousand-Character Text and Chinese radicals, to Roget's thesaurus. According to the comparison of the Thousand-Character Text and Roget's thesaurus, most of the 39 sections of Roget's thesaurus are relevant to Chinese characters in the Thousand-Character Text. The correlation coefficient is around 0.90. In the case of Chinese radicals, 30 sections of Roget's thesaurus are relevant to the radicals. The correlation coefficient is around 0.85, showing considerable relevancy between Chinese radicals and sections of Roget's thesaurus, as well.

Keywords: Thousand-Character Text, Chinese Radicals, Roget's Thesaurus, Ontology, Interontologia.

1 Introduction

With the development of the Internet, the volume of information is now incomparable with that in the past. In order to manage such a large amount of information, we need standardized classification systems. A standardized classification system can be created based on human cognitive ability to classify things. Everything in the world has its own unique characteristics, by which it is classified into specific categories, and we understand things more easily by associating them to related categories. In this way, we simplify information processing and understand perceived things better by classifying them systematically according to their characteristic.

The lexical classification system covered in this study can be divided into various types according to the use of words or information. Examples of application are in the areas of artificial intelligence, computational linguistics and information communication include information search, knowledge management, information system design, ontology building, machine translation, and dictionary compilation. There are also implemented lexical classification systems related to the vocabulary resources of natural languages such as Roget's thesaurus [1], WordNet [2], Lexical FreeNet [3], Kadokawa thesaurus [4], and EDR [5]. Cases of ontology building include KR Ontology [6], CYC Ontology [7], Mikrokosmos Ontology[8], SENSUS Ontology [9] and

* Corresponding author.

B.-H. Kang and D. Richards (Eds.): PKAW 2010, LNAI 6232, pp. 126–134, 2010.

HowNet[10], and there are business applications of ontology such as Enterprise Ontology[11], UMLS[12], UNSPSC[13], RosettaNet[14], ISO 2788[15] and ANSI Z39.19[16].

Lexical classification is concept classification by nature. Lexical classification systems mentioned above suggest that there are various concept classification systems today. It is said that people have the same cognition, memory, causal analysis, categorization, and reasoning process. They assume that if there is any difference, it is not from difference in cognitive process but from difference in culture or education [17]. As mentioned above, in the current situation that various concept classification systems are being used in different application areas, it is keenly required to interlock concept classification systems and intelligent information systems. In response to the demand, research is being made on ontology mapping, merge and integration, and semantic integration [18, 19]. A main research method is the utilization of shared ontology or finding mapping in ontological features.

However, if there is a general concept classification system (*interontologia*) [20] as a reference classification system, through which it will become more systematic and easier to integrate concept classification systems semantically. Thus, as a case study on general concept classification system, the present study examines the relevancy of lexical categorization between the Thousand-Character Text [21, 22], which is a representative Eastern classic, and Roget's thesaurus[1], which is a famous Western classified lexicon. In addition to this, we also investigate the relevancy between Chinese radicals and the thesaurus. Through this study, we analyze similarities between the two in categorization and classification.

2 Lexical Ontologies: The Thousand-Character Text, Chinese Radicals, and Roget's Thesaurus

The Thousand-Character Text (千字文) was written by Zhou Xingsi (周興嗣) by order of Emperor Wu (武帝) in the Liang (梁) Dynasty of China in around the 6th century, and transmitted and distributed to Korea early in ancient times. This is a representative classical Chinese textbook used widely to teach children. The oldest Thousand-Character Text annotated with pronunciation and meaning in Korean is the version of Han Seok-Bong published in 1583. There is also a record on an earlier version published in Gwangju in 1575. The Thousand-Character Text is old four-character verse composed of a total of 250 four-character phrases or 125 couplets and its materials are Chinese history, culture, etc. [21, 22].

Roget's thesaurus [1] was first published in 1852 by English surgeon Peter Mark Roget. This is the first synonym/antonym dictionary. Roget's thesaurus is not in meaningless alphabetical order. The thesaurus classifies lexical knowledge systematically. The top hierarchy is composed of 6 classes, under which are divisions. Each division is again subdivided into sections. In each hierarchy is unique entry information, and at the end of the hierarchical structure are listed a total of 1044 categories. Each category has a list of synonyms by part of speech. On the other hand, if a specific word in the list of synonyms refers to another category, the reference is expressed in the form of "Vocabulary &c. (Entry word) Entry number."

There is a study on lexical classification systems in Korean representative classics such as the Thousand-Character Text, Yuhap (類合) and Hunmongjahoi(訓蒙字會) [22]. In the study, they argued that the Thousand-Character Text is structured well and has a clear system. In addition, they emphasized the accuracy of its classification system that does not allow even a repetition of the same character among the 1000 characters. Largely according to semantic paragraph (content), they classified Thousand-Character Text as follows: astronomy, nature, royal task, moral training, loyalty and filial piety, virtuous conducts, five moral disciplines, humanity and justice, palace, meritorious retainers, feudal lords, topography, agriculture, mathematics, quiet life, comfort, miscellaneous affairs, skills, admonition, etc. They concluded that in presenting Chinese characters by semantic paragraph, the Thousand-Character Text arranges basic Chinese characters appropriately and is outstanding in terms of lexical system and the perception of basic Chinese characters.

Chinese radicals are index keys or classifiers which are used for organizing entries in Chinese dictionary. Hàn dynasty(漢朝) scholar Xǔ Shèn(許慎) categorized all the Chinese characters with a system of 540 graphic elements that was called bùshǒu (部首). The elements are character components and denote some common semantic or phonetic characteristics. Chinese lexicographers had continued to refine this system for indexing Chinese characters. The number of Chinese radicals was reduced to 214 in the dictionary Zìhuì(字彙) in 1615. The set of radicals became standard and is still used in Chinese dictionaries today [23].

3 Relevancy Analysis

This study has conducted analysis on concept relevancy through 5 steps as follows.

Step 1. Building Master Databases: In this step, we sort out Chinese characters from the Thousand-Character Text, and words from Roget's thesaurus. Then build the master databases for the characters in the Thousand-Character Text, for the radicals in Chinese Radicals, and for the words in Roget's thesaurus, respectively.

Step 2. Identifying Meanings of Chinese Characters in English: In this step, we translate the meaning(s) of each Chinese character in the Thousand-Character Text and Chinese radicals into English words. Then we keep the translation in a database, referring to Chinese-English dictionary Kingsoft2008[24], Classical Chinese Character Frequency List[25], YellowBridge[26], and CHINAKNOWLEDGE[27].

Step 3. Filling Fields in Master Databases with English Equivalents: In this step, we determine an English equivalent for each Chinese character in the Thousand-Character Text and for each Chinese radical. And then fill the fields with the words in corresponding master databases.

Step 4. Mapping Chinese Characters and Radicals to Roget's Thesaurus Categories: In this step, we map English words for Chinese characters in the Thousand-Character Text and for Chinese radicals, into categories of Roget's thesaurus. And then keep the category information in the prearranged fields in the master databases.

Step 5. Analyzing Mapping Results: Finally, we analyze each relevancy of Thousand-Character Text and Chinese Radicals with respect to Roget's thesaurus using graphs and correlation coefficients based on the mapping data.

4 Analysis Results

4.1 The Thousand-Character Text and Roget's Thesaurus

Among the 1,044 categories of Roget's thesaurus, 424 categories have one or more corresponding Chinese characters while 620 categories do not have any corresponding Chinese characters. We also have analyzed the mapping on the section level of Roget's thesaurus, which are higher categories in the hierarchy of the thesaurus. Table 1 compares the mapping on the category level with the one on the section level of Roget's thesaurus.

Table 1. Results of mapping by classification level (Chinese characters of Thousand-Character Text)

Level	Number of mapped entries	Number of unmapped entries	Total	Mapping rate (%)
Roget's Category	424	620	1,044	41
Roget's Section	38	1	39	97

Fig. 1 shows the correspondence between Chinese characters in the Thousand-Character Text and Roget's thesaurus on the section level. We can see that the number of Chinese characters changes as the number of Roget's thesaurus categories does. We have obtained the correlation coefficient r_{xy} for the association between the number of Chinese characters in the Thousand-Character Text and Roget's thesaurus categories on the section level.

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}} \tag{1}$$

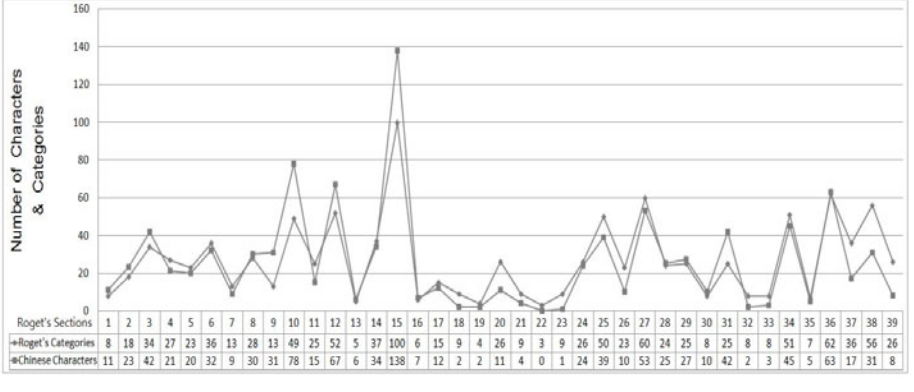

Fig. 1. Correspondence between the Thousand-Character Text and Roget's thesaurus on the section level

In the correlation analysis, the number of Roget categories is defined as X, the number of mapped Chinese characters as Y, and the number of Roget sections as N. For Equation (1), the values of variables are as follows.

$$\sum X_i = 1044,\ \sum Y_i = 1000,\ \sum X_i Y_i = 46038,\ \sum X_i^2 = 44474,\ \sum Y_i^2 = 53164$$

If these values are substituted for the variables in Equation (1), we obtain $r_{xy} = 0.90$, showing quite a high correlation between the Thousand-Character Text and Roget's thesaurus on the section level.

4.2 The Chinese Radicals and Roget's Thesaurus

We have also analyzed the mapping results between Chinese radicals and sections of the Roget's thesaurus. Table 2 shows the mapping rate on the section level of Roget's thesaurus. Table 3 shows the results. Among a total of 39 sections in Roget's thesaurus, three do not have any corresponding radicals, but 36 sections have one or more. The correlation coefficient is as high as 85%.

Table 2. Results of mapping by classification level (Chinese radicals)

Level	Number of mapped entries	Number of unmapped entries	Total	Mapping rate (%)
Roget's Sections	36	3	39	92

Fig. 2 and Table 3 shows the correspondence between Chinese radicals and Roget's thesaurus on the section level. We can see that the number of Chinese radicals changes as the number of Roget's thesaurus categories does.

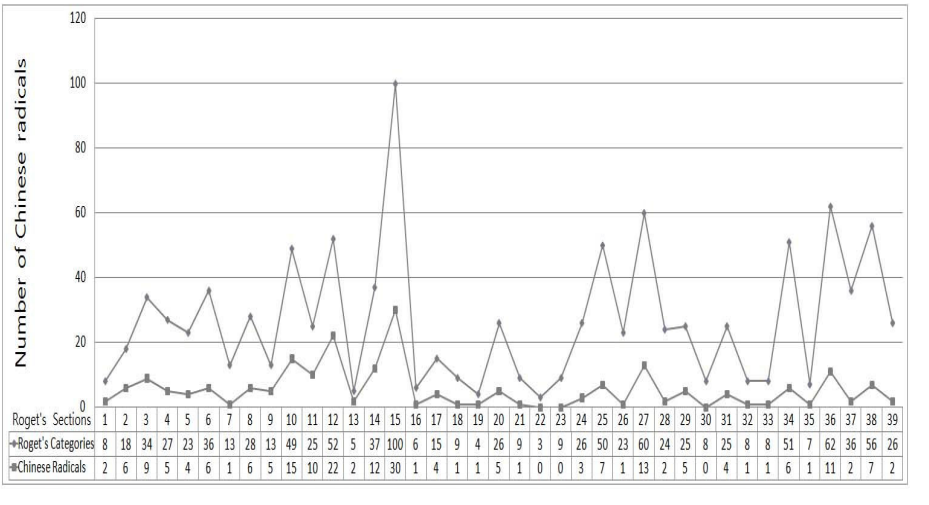

Fig. 2. Correspondence between Chinese radicals and Roget's thesaurus on the section level

In the correlation analysis, the number of Roget categories is defined as *X*, the number of mapped Chinese radicals as *Y*, and the number of Roget sections as *N*. For Equation (1), the values of variables are as follows.

$$\sum X_i = 1044,\ \sum Y_i = 214,\ \sum X_i Y_i = 9951,\ \sum X_i^2 = 44474,\ \sum Y_i^2 = 2652$$

If these values are substituted for the variables in Equation (1), we obtain $r_{xy} = 0.85$, showing a relatively high correlation between Chinese radicals and Roget's thesaurus on the section level.

Table 3. Mapping from Chinese radicals to Roget's thesaurus on the section level

No	Class	Roget's Sections	Chinese Radicals	Count
01	1	existence	立, 身	2
02	1	relation	韋, 方, 己, 比, 自, 血	6
03	1	quantity	寸, 丿, 大, 小, 尢, 支, 片, 而, 頁	9
04	1	order	鬥, 冂, 丨, 乙, 氏	5
05	1	number	鼎, 二, 又, 疋	4
06	1	time	夕, 子, 幺, 老, 辰, 長	6
07	1	change	艮	1
08	1	causation	力, 父, 虍, 龍, 屮, 爿	6
09	2	space in general	凵, 匚, 斗, 臼, 鬲	5
10	2	dimensions	亠, 勹, 匸, 厂, 口, 宀, 巾, 广, 毛, 瓦, 衣, 貝, 足, 高, 髟	15
11	2	form	丿, 刀, 彐, 彡, 穴, 角, 谷, 豆, 門, 齊	10
12	2	motion	儿, 入, 口, 夂, 戶, 瓜, 肉, 至, 舟, 行, 襾, 走, 辵, 酉, 飛, 食, 鬯, 魚, 黍, 皿, 缶, 癶	22
13	3	matter in general	山, 月	2
14	3	inorganic matter	土, 川, 气, 水, 田, 米, 糸, 雨, 革, 風, 骨, 麥	12
15	3	organic matter	一, 丶, 尸, 歹, 火, 甘, 白, 目, 羊, 耳, 色, 虫, 豕, 赤, 辛, 隹, 面, 韭, 音, 香, 馬, 鳥, 鹵, 鹿, 黃, 黑, 黽, 鼓, 龜, 龠	30
16	4	operations of intellect in general	示	1
17	4	precursory conditions and operations	几, 見, 鼻, 釆	4

Table 3. (*continued*)

No	Class	Roget's Sections	Chinese Radicals	Count
18	4	materials for reasoning	耒	1
19	4	reasoning processes	舛	1
20	4	results of reasoning	玄, 首, 艸, 士, 青	5
21	4	extension of thought	爻	1
22	4	creative thought		0
23	4	nature of ideas communicated		0
24	4	modes of communication	厶, 无, 毋	3
25	4	means of communicating ideas	丶, 文, 曰, 聿, 舌, 言	6
26	5	volition in general	鼠	1
27	5	prospective volition	冖, 匕, 工, 廾, 弋, 斤, 欠, 牙, 用, 疒, 石, 肉, 金	13
28	5	voluntary action	生, 彳	2
29	5	antagonism	戈, 殳, 矛, 矢, 皐	5
30	5	results of voluntary action		0
31	5	general intersocial volition	臣, 隶, 邑, 里	4
32	5	special intersocial volition	止	1
33	5	conditional intersocial volition	卩	1
34	5	possessive relations	皮, 网, 車, 攴, 手, 爪	6
35	6	affections in general	禾	1
36	6	personal affections	夂, 夊, 八, 十, 日, 牛, 人, 玉, 羽, 豸, 辮, 齒	12
37	6	sympathetic affections	弓, 心	2
38	6	moral affections	麻, 干, 非, 女, 木, 竹, 犬	7
39	6	religious affections	卜, 鬼	2

5 Conclusions and Future Research

The present study has examined concept relevancy between the Thousand-Character Text and Roget's thesaurus. Moreover, we have also conducted analysis on concept relevancy between Chinese Radicals and Roget's thesaurus. From the result of our experiment, we may say that there is an *interontologia* behind Roget's thesaurus and the Thousand-Character Text, or Chinese radicals.

Tasks for future research include: (1) complementing omitted parts in mapping of Chinese characters in the Thousand-Character Text to Roget's thesaurus categories with the 1800 commonly used Chinese characters in Korea and comparing the results; and (2) analyzing difference between comparison of the Thousand-Character Text and Roget's thesaurus on the category and section levels and (3) trimming 214 Chinese radicals into 100. From these studies, we can expect to have a set of Chinese characters for a refined lexical knowledge classification system. Lastly, based on the character set, we will develop a new lexical category system applicable to knowledge classification.

Acknowledgments. This work was supported by the Korea Research Foundation Grant funded by the Korean Government. (KRF-2008-313-H00009)

References

1. Roget's Thesauri, http://www.bartleby.com/thesauri/
2. WordNet, http://wordnet.princeton.edu/
3. Lexical FreeNet, http://www.cinfn.com/doc/
4. Ohno, S., Hamanishi, M.: New Synonyms Dictionary, Kadogawa Shoten, Tokyo (1981) (Written in Japanese)
5. The EDR Electronic Dictionary, http://www2.nict.go.jp/r/r312/EDR/index.html
6. KR Ontology, http://www.jfsowa.com/ontology/
7. CYC Ontology, http://www.cyc.com/
8. Mikrokosmos Ontology, http://crl.nmsu.edu/Research/Projects/mikro/htmls/ontology-htmls/onto.index.html
9. SENSUS Ontology, http://www.isi.edu/natural-language/projects/ONTOLOGIES.html
10. HowNet, http://www.keenage.com/html/e_index.html
11. Enterprise Ontology, http://www.aiai.ed.ac.uk/project/enterprise/enterprise/ontology.html
12. UMLS, http://www.nlm.nih.gov/research/umls/
13. UNSPSC, http://www.unspsc.org/
14. RosettaNet, http://www.rosettanet.org
15. ISO 2788, http://www.collectionscanada.gc.ca/iso/tc46sc9/standard/2788e.htm
16. ANSI Z39.19, http://www.niso.org/standards/resources/Z39-19-2005.pdf
17. Nisbett, R.E.: The Geography of Thought: How Asians and Westerners Think Differently... and Why. Simon & Schuster, New York (2004)
18. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. The Knowledge Engineering Review 18(1), 1–31 (2003)
19. Noy, N.F.: Semantic Integration: A Survey of Ontology-Based Approaches. SIGMOD Record 33(4), 65–70 (2004)

20. Kim, S.-R., Yang, J.-G., Bae, J.-H.J.: An Investigation of an Interontologia: Comparison of the Thousand-Character Text and Roget's Thesaurus. In: Li, W., Mollá-Aliod, D. (eds.) ICCPOL 2009. LNCS (LNAI), vol. 5459, pp. 394–401. Springer, Heidelberg (2009)
21. Kim, J.-T., Song, C.-S.: Comparison of Vocabulary Classification Systems among Thousand-Character Text, Yuhap, and Hunmongjahoi, Korean Literature Society, Linguistics and Literature, vol. 52, pp. 159–192 (1991) (written in Korean)
22. Jin, T.-H.: Problems in the Translations and Sounds of Thousand-Character Text. Hangeul-Chinese Character Culture 104, 80–82 (2008) (written in Korean)
23. Wikipedia: Section headers of a Chinese dictionary, http://en.wikipedia.org/wiki/Section_headers_of_a_Chinese_dictionary
24. Kingsoft2008 (谷歌金山词霸), http://g.iciba.com/
25. Classical Chinese Character FrequencyList, http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=CL
26. YellowBridge, http://www.yellowbridge.com
27. CHINAKNOWLEDGE, http://www.chinaknowledge.de/Literature/radicals.html