A Muti-Resolution Approach to Restaurant Named Entity Recognition in Korean Web

Bo-Yeong Kang[1] and Dae-Won Kim[2*]

[1] **School of Mechanical Engineering, Kyungpook National University, Daegu 702-701, Korea**
[2] **School of Computer Science and Engineering, Chung-Ang University, Seoul 156-756, Korea**

Abstract

Named entity recognition (NER) technique can play a crucial role in extracting information from the web. While NER systems with relatively high performances have been developed based on careful manipulation of terms with a statistical model, term mismatches often degrade the performance of such systems because the strings of all the candidate entities are not known a priori. Despite the importance of lexical-level term mismatches for NER systems, however, most NER approaches developed to date utilize only the term string itself and simple term-level features, and do not exploit the semantic features of terms which can handle the variations of terms effectively. As a solution to this problem, here we propose to match the semantic concepts of term units in restaurant named entities (NEs), where these units are automatically generated from multiple resolutions of a semantic tree. As a test experiment, we applied our restaurant NER scheme to 49,153 nouns in Korean restaurant web pages. Our scheme achieved an average accuracy of 87.89% when applied to test data, which was considerably better than the 78.70% accuracy obtained using the baseline system.

Key Words: Named entity classification, semantic feature, multi-resolution approach

1. Introduction

Automatic recognition of named entities such as the names of persons, organizations, locations, expressions of times has emerged as a popular technique for information extraction and knowledge management on the web. Traditional approaches to named entity recognition (NER) typically describe the text with a single representation composed of terms and simple term-level features (initial capitalization and character type such as Hiragana or Katakana) in the text; they attempted to locate terms or phrases in raw text and classify them into predefined categories (persons, organizations). These NER systems can achieve relatively high performance through careful manipulation of terms with a statistical model; in such approaches, however, term mismatches can deteriorate the performance because the strings of all the candidate entities are not known a priori to a system. Therefore, in the present study we sought to develop a new method for restaurant NER based on a multi-resolution approach, which uses semantic features of terms at multiple semantic resolutions, and can therefore effectively handle the variations of terms.

2. Related Work

Korean NER systems have typically been based on rule-based approaches using lexicons and grammars; however, these methods require the creation of grammars and lexicons, making them labor-intensive and time-consuming [1, 2, 3]. Some researchers attempted to develop statistical NER methods based on SVMs and conditional random fields (CRFs) to find NEs for people, locations, and organizations in Korean [4, 5]. However, these previous works have the following shortcomings. First, in English and Japanese NER, various simple term-level features such as initCap and character type can be used as good clues for recognition; however, these simple term-level features rarely occur in Korean, making it difficult to exploit these features in Korean NER. Second, previous approaches were limited to utilizing the term string itself and simple term-level features, and did not exploit the semantic features of terms which can handle the variations of terms effectively. To address the problem of lexical-level term mismatches, here we attempt to handle the semantic content of term units in named entities (NEs) in the restaurant domain. In our proposed approach, when an NER system identifies a candidate entity that is likely to be a restaurant

NE, for example, the decision should be made at the semantic rather than the lexical level. The following example shows a representative case:

Train :

Ruth's Chris Steak House

Lucky Dragon Chinese Restaurant

Empire Turkish Grill

Test :

California Pizza Kitchen

Hot and Hot Fish Club

It is clear that the relationship between the terms in the training and test sets plays a crucial role in recognizing the restaurant NE of California Pizza Kitchen and Hot and Hot Fish Club in the test set. However, the relationship cannot be shown easily at a lexical level. As a solution to this problem, we propose matching the semantic concepts automatically generated from multiple resolutions of a semantic tree. A semantic tree is a tree of concepts (nodes) interlinked by semantic and lexical relationships, and the resolution refers to the level (depth) of a concept in the semantic tree. Concepts at low (coarse) resolution in the semantic tree are more generic or have broader meanings than those at high (fine) resolution. Once the low resolution concepts are determined by analyzing the set of training NEs (e.g., FOOD or BUILDING), NEs with the same or related concepts in the test set can be selected as candidates for the restaurant NER as follows:

Train :

Ruth's Chris::PERSON

Steak::FOOD

House::BUILDING

Test :

California::LOCATION

Pizza::FOOD

Kitchen::BUILDING

3. Restaurant Named Entity Recognition

Web page content in the English language restaurant domain typically consists of restaurant NEs such as menus and addresses. A corpus of restaurant web pages revealed the following characteristics: (1) Web page source codes are composed of HTML tags intermingled with free text, hence, the reading order of the free text after removing HTML tags is often different from that of the original layout in the web page. (2) Since HTML layout information changes dynamically, it is almost impossible to apply the

layout information from one web site to web pages supported by another web site. (3) In web pages, restaurant NEs are frequently used alone as one sentence without any contextual information. Given these characteristics of web pages, and their effect on extracting contextual information, we utilize only internal evidence to recognize the NEs without the use of external evidence.

The presentation style and contents of the Korean web page are similar to those of English language restaurant web pages. However, while an English language restaurant NE is represented by a phrase type composed of several terms and space characters (i.e. Leesangman dish house), a Korean language restaurant NE is represented by a single multi-word type composed of several terms without space characters to segment the terms [9] (i.e., 이상만요리집((Leesangmandishhouse)). The following shows a few sample NEs extracted from [22]: here, Leesangman is the name of a person, and Gabdong and Mapo are location names. For easy comprehension of the samples, hypens (-) and uppercase are used to separate the terms and to indicate the names of a person, location, and organization. In original Korean NEs, hypens and uppercase characters are not used:

이상만-요리-집(Leesangman-dish-house)
원조-마포-숯불-갈비(original-Mapo-charcoal fire-rib)

Among 1,010 Korean restaurant NEs that we examined in Menupan.com [8], 95.6% were of single multi-word type and 4.4% of phrase type. Given these characteristics of Korean restaurant NEs in web pages, we decided to develop the preprocessing module to handle single multi-word type NEs efficiently for restaurant NER.

3.1 Multi-Word Decomposition

After the target web pages for restaurant NER have been crawled, all the crawled pages are preprocessed by an HTML parser and a Korean POS (Part Of Speech) tagger, KLT [10], to split sentences and extract nouns. All the nouns extracted by KLT are regarded as the candidate restaurant NEs. Based on the examination that most of Korean restaurant names are of single multi-word type as mentioned above, the system decomposes the candidate NE into several term units to utilize internal evidence of the candidate restaurant NEs. For multi-word decomposition, we still employ KLT. The following shows an example:

NE: 할매-곰탕-집(grandmother-thick broth-house)

KLT: [할매(grandmother), 곰탕-집(thick broth-house)]

Postprocessing: [할매(grandmother), 곰탕(thick broth), 집(house)]

[할매(grandmother), 곰탕-집(thick broth-house)] is a KLT decomposition result of a restaurant NE. Here, we
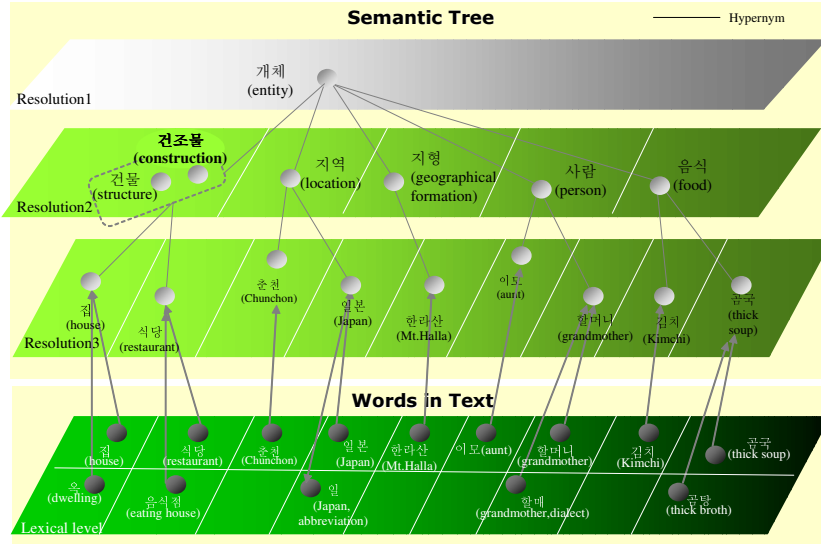
Figure 1: Semantic tree of concepts

can see that the one-syllable noun, 집(house), was not decomposed from 곰탕-집(thick broth-house) because KLT does not allow one-syllable nouns to be detached independently. However, since one-syllable nouns can be important features for restaurant NER, our system decomposes 곰탕-집(thick broth-house) into 곰탕(thick broth) and 집(house), if 곰탕-집(thick broth-house) is not registered as an entry in the decomposition dictionary.

3.2 Multi-Resolution Modeling

When applied to the decomposed multi-word α, the transformation function ϕ generates the resolution-based concept vector γ_i for each term $\alpha_i \in \alpha$ based on the semantic tree Δ. Therefore, we define the multi-resolution model, in which the semantics of terms can be captured in a wide range of levels of detail and the terms can be transformed into the concepts at any one of those resolutions on demand as follows:

Definition 3.1. A multi-resolution model for NEs is defined as a model in which the semantics of a term can be represented in a wide range of levels of detail and a term can be reconstructed as concepts at any one of those resolutions on demand. It is composed of $< \alpha, \gamma, \Delta, \Phi >$ where α is a vector of input terms, γ the vector of output concepts reconfigured from α, and Δ a semantic tree of concepts. A transformation function, $\Phi : \alpha_i \in \alpha \rightarrow \gamma_i$, transforms an input term in $\alpha_i \in \alpha$ into a concept vector, γ_i, at a certain semantic resolution of Δ.

As the criterion for the multi-resolution representation of terms, we referenced the semantic tree of concepts from

Korean WordNet [11, 18] as depicted in Fig. 1. Korean WordNet is a large lexical database in Korean, where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Then, to denote each concept, we can use a preferred term which refers to the synset in which the concept resides.

In Fig. 1, the set of synonyms comprised of *structure* and *construction* represents one concept referring to an object that is constructed; hence this concept is denoted by the preferred term *construction*. The concepts in the semantic tree are interlinked by means of conceptual-semantic and lexical relations. Here, the resolution indicates the level (depth) of a concept in the tree. Concepts with high resolution can be generalized to more abstract concepts with low resolution by hypernym relations. For instance, the low-resolution concept *entity* has a more generic or broad meaning than the high-resolution concept *restaurant*.

Table 1 shows how a decomposed restaurant NE, [할매(grandmother)-곰탕(thick broth)-집(house)], can be converted into concepts at various resolutions. For instance, the possible concepts of [할매(grandmother)-곰탕(thick broth)-집(house)] are assigned as [사람(person)-음식(food)-구조물(construction)] at resolution 2.

However, since a semantic tree is not a perfect tree with all leaf concepts at the same resolution, the resolution structure for concepts depends on the criteria used to decide the resolution, as shown in Figs. 2 and 3. Fig. 2 shows the semantic tree in which the resolution is based on and starts from the root concept, *entity*, while Fig. 3 depicts the tree based on the leaf concepts, *house*,
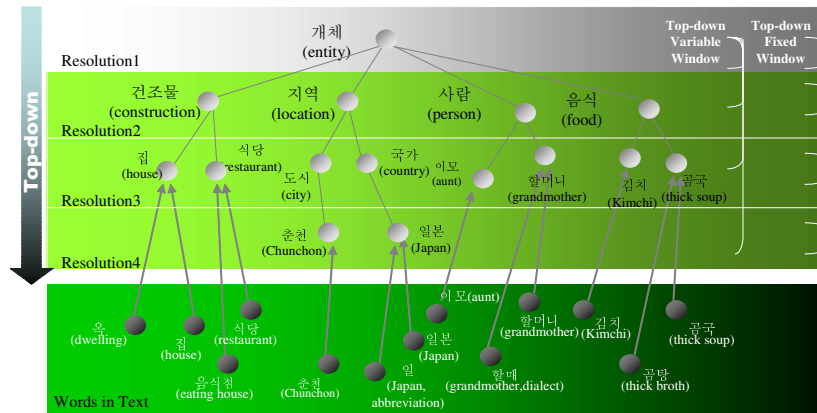
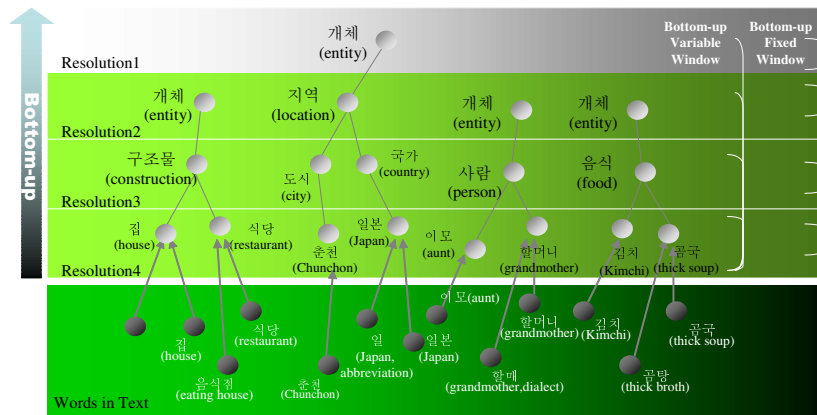Figure 2: Semantic tree based on root concept (top-down)

Figure 3: Semantic tree based on leaf concepts (bottom-up)

restaurant, *grandmother*, and so on. The concepts *house* and *restaurant* are at resolution 3 in Fig. 2, but at resolution 4 in Fig. 3.

To handle concepts at various resolutions in the tree, we developed top-down and bottom-up transformation functions that convert the terms in the restaurant NE into the concepts at each resolution of the tree. The top-down function obtains the concepts at a certain resolution referring to the resolution structure based on the root concept, while the bottom-up function refers to the resolution structure based on leaf concepts. Then, each transformation function uses the notion of a window, the span of resolutions, that the system should look within to obtain concepts at a certain resolution. A variable window (v) indicates a variable span of resolutions, and a fixed window (f) indicates a fixed span of resolutions. Therefore, for each traverse type in the tree, we developed four types of transformation functions ϕ:

$\phi_{top,r,v}$ maps a term into concepts found from the root concept to resolution (r) with (v).

$\phi_{top,r,f}$ maps a term into a single concept found at resolution (r) with (f).

$\phi_{bottom,r,v}$ maps a term into concepts found from the leaf concepts to (r) with (v).

$\phi_{bottom,r,f}$] maps a term into a single concept found at resolution (r) with (f).

Table 2 shows how the concepts at each resolution can be assigned using the four transformation functions for the decomposed restaurant NE based on Figs. 2 and 3. For example, the concepts of the decomposed restaurant NE using $\phi_{top,r=2,v=2}$ are assigned as [[개체(entity), 사람(person)], [개체(entity), 사람(food)], [개체(entity), 구조물(construction)]], which corresponds to the vector of concepts from the root concept resolution to the resolution ($r = 2$) with variable window size ($v = 2$). The function $\phi_{bottom,r=2,v=3}$ obtains the vector of concepts from the leaf concept resolution to the resolution

Table 1: Concept vectors from 4 transformation functions for [할매(grandmother)-곰탕(thick broth)-집(house)]

	α_1	α_2	α_3
NE	할매(grandmother)	곰탕(thick broth)	집(house)
	γ_1	γ_2	γ_3
$\Phi_{top,r=2,v=2}$	개체(entity)	개체(entity)	개체(entity)
	사람(person)	음식(food)	구조물(construction)
$\Phi_{bottom,r=2,v=3}$	할매(grandmother)	곰국(thick soup)	집(house)
	사람(person)	음식(food)	구조물(construction)
	개체(entity)	개체(entity)	개체(entity)
$\Phi_{top,r=2,f=1}$	사람(person)	음식(food)	구조물(construction)
$\Phi_{bottom,r=2,f=1}$	개체(entity)	개체(entity)	개체(entity)

($r = 2$) with variable window size ($v = 3$), therefore, it results in [[할매(grandmother), 사람(person), 개체(entity)], [곰국(thick soup), 음식(food), 개체(entity)], [집(house), 구조물(construction), 개체(entity)]]. The function $\phi_{top,r=2,f=1}$ indicates the concepts at the resolution ($r = 2$) based on the root concept with fixed window ($f = 1$), and the final function $\phi_{bottom,r=2,f=1}$ obtains concepts at the resolution ($r = 2$) based on the leaf concepts with fixed window ($f = 1$).

4. Experimental Results

4.1 Experimental settings

To test the efficacy of the proposed NE extraction method, we conducted a classification experiment using nouns extracted from Korean language restaurant domain web pages. For training data construction, 467 web pages were crawled using WIRE (Web Information Retrieval Environment) [12], focusing on the popular Korean restaurant web site Joyfood.com [13]. All the web pages in the crawled data were preprocessed by an HTML parser [14] and KLT [10] of a Korean POS (Part Of Speech) tagger to extract nouns. As the reference of the semantic tree for multi-resolution representation, we used Korean WordNet, KorLex 1.5 [18]. KorLex 1.5 was developed from 2004 to 2007, which contains about 130,000 synonyms (synsets) and 150,000 word senses for nouns, verbs, adjectives, adverbs, and classifiers.

The proposed method was applied to a set of 49,153 nouns extracted from the preprocessed web pages. This set of nouns contained a positive set of 1,208 manually tagged restaurant NEs (2.5%) for restaurant NE class and a negative set of 47,945 restaurant relevant nouns (97.5%) such as names of dishes, people, beverages, and foodstuffs, as well as phone numbers, money-related nouns, and so on. The performance of the proposed method was evaluated using 10-fold cross validation on the extracted data. Given the resolution-based features generated from an input restaurant NE, we used the support vector machines (SVMs) to classify the restaurant NEs.

For classification on the imbalanced dataset composed of 2.5% positive and 97.5% negative, the standard accuracy can lead to incorrect results since it is strongly dominated by the majority class [16]. Widely used measures to evaluate the system performance on imbalanced data are true positive rate, true negative rate, and Geometric-mean [16]. The G-mean measures the balanced performance of a learning algorithm between positive and negative classes.

To the best of our knowledge, there have been few reports on the performance of restaurant NER method. Therefore, we decided to use the SVM based NER method by Isozaki and Kazawa [17] as the baseline, which considers all the term features obtainable from the multi-word decomposition by KLT.

4.2 Performance Evaluation

To examine the effectiveness with which the proposed method recognizes the restaurant NEs of the test set, we ran NER experiments on a sequence of transformation functions as follows: (1) $\phi_{top,r=1..4,v=1..14}$, which obtains the vector of concepts from the root concept resolution to the resolution ($r = 1..14$) with variable window size ($v = 1..14$); (2) $\phi_{top,r=1..14,f=1}$ indicates the concepts at the resolution ($r = 1..14$) based on the root concept with fixed window ($f = 1$); (3) $\phi_{bottom,r=14..1,v=1..14}$ obtains the vector of concepts from the leaf concept resolution to the resolution ($r = 14..1$) with variable window size ($v = 1..14$); (4) the final function $\phi_{bottom,r=14..1,f=1}$, obtains concepts at the resolution ($r = 14..1$) based on leaf concepts with fixed window ($f = 1$). Fig. 4 and Fig. 5 show the overall results for the test set, expressed in terms of the G-mean, for the top-down and bottom-up functions with varying windows. From Fig. 4, we can see that $\phi_{top,r=1..4,f=1}$ showed better performance than
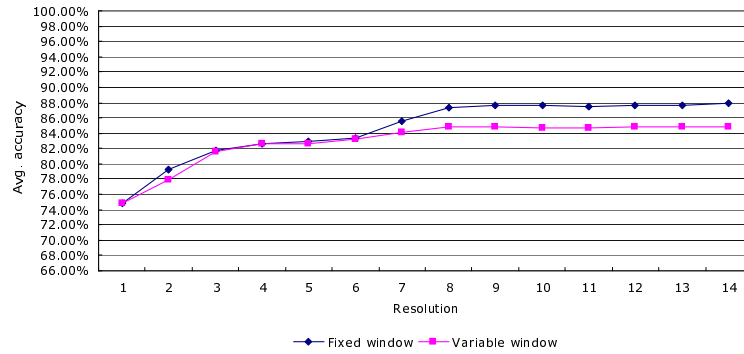
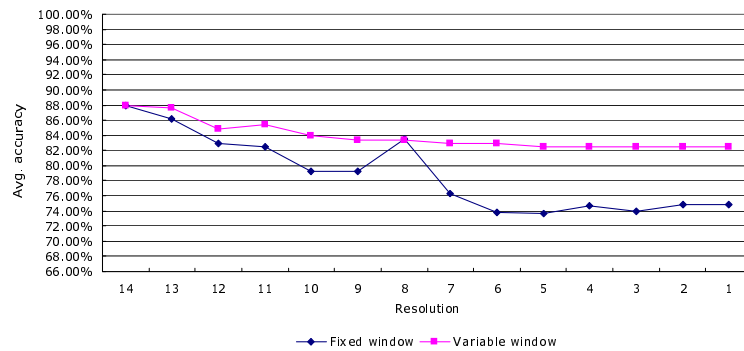Figure 4: G-mean on top-down functions with v and f

Figure 5: G-mean on bottom-up functions with v and f

$\phi_{top,r=1..4,v=1..4}$ in the overall categories, and that the best G-mean result (0.873) was obtained using the fixed window, $\phi_{top,r=14,f=1}$.

To examine in greater detail the performance of the function $\phi_{top,r=1..4,f=1}$, Table 3 lists the performance of this function for positive and negative classes, and the G-mean performance results for the test set. From Table 3, we can see that the best performance (0.98 and 0.77 for the negative and positive classes of the test set respectively) was obtained in the final resolution 14. The results also show that the G-mean of the test set improved remarkably on going from resolution 1 to 8, but then showed little improvement up to resolution 14. These findings indicate that the proposed method can efficiently handle the term variations in the test set from resolution 8, although it showed the best performance in the final resolution.

Unlike the top-down transformation, the bottom-up with variable window, $\phi_{bottom,r=14..1,v=1..4}$, showed better performance than the function with a fixed window, $\phi_{bottom,r=14..1,f=1}$, as shown in Fig. 5. The detailed results of $\phi_{bottom,r=14..1,v=1..4}$, are listed in Table 4. From the results, we see that among the bottom-up transforma-

Table 2: Performance at $\Phi_{top,r=1...14,f=1}$

r	TNR	TPR	G-mean	r	TNR	TPR	G-mean
1	0.56	0.93	0.72	8	0.97	0.77	0.86
2	0.71	0.87	0.78	9	0.98	0.77	0.87
3	0.86	0.76	0.81	10	0.98	0.77	0.86
4	0.93	0.71	0.82	11	0.98	0.77	0.86
5	0.93	0.72	0.82	12	0.98	0.77	0.86
6	0.96	0.69	0.82	13	0.98	0.77	0.87
7	0.97	0.73	0.84	14	0.98	0.77	0.87

tion functions, the best G-mean result (0.87) was obtained in the final resolution ($r = 14$). As the resolution was decreased, the G-mean of the fixed window approach deteriorated much more rapidly than that of the variable window. This finding indicates that the concept features in the final resolution can play an important role in the NER performance; thus the variable window approach with concept features in the final resolution shows more stable performance than the fixed window.

Table 3: Performance at $\Phi_{bottom,r=14...1,v=1...14}$

r	TNR	TPR	G-mean	r	TNR	TPR	G-mean
14	0.98	0.77	0.87	7	0.98	0.71	0.83
13	0.97	0.77	0.87	6	0.98	0.71	0.83
12	0.98	0.71	0.83	5	0.98	0.71	0.83
11	0.98	0.71	0.84	4	0.98	0.71	0.83
10	0.98	0.71	0.83	3	0.98	0.71	0.83
9	0.98	0.71	0.83	2	0.98	0.71	0.83
8	0.98	0.71	0.83	1	0.98	0.71	0.83

Table 4: Performance over the baseline

	TNR	TPR	G-mean
Baseline	0.95	0.62	0.77
Proposed	0.98	0.77	0.87

Table 5 shows the performance results obtained using the baseline system and the proposed method with the optimal transformation functions of $\phi_{top,r=14,f=1}$ and $\phi_{bottom,r=14,v=1}$. Compared to the baseline system, the proposed method shows significantly higher overall performance. For the incoming stream of 4,900 nouns from the test set, the performance improvement of the proposed method over the baseline method was 15.3% for the positive class and 3.1% for the negative class. On average, the proposed method improved the performance 10.3% over the baseline.

4.3 Error Analysis

We analyzed the NER errors of the proposed method for one set from among the 10 training and test sets that were used for the cross validation. Since the error rate of the positive class in the test set was relatively high compared to that of the negative class, error analysis was conducted on the positive class in the test set.

Our error analysis of the results of the proposed method revealed that many errors stemmed from insufficient training data and semantic tree entries; 9.12% of the total number of errors came from the fact that the training data and the semantic tree did not support the terms that appeared in the restaurant NEs. For instance, our system then classifies the sample data [신길(Singil(town)), 장어(eel)] as the negative class based on the concept food of 장어(eel), because 신길(Singil(town)) was not in the training data and even in the semantic tree.

Another issue that we should consider in the proposed method is the problem of ambiguity, which accounted for 7.10% of the total errors. In the present work, we took the first sense of a term among several senses listed in the Korean WordNet index file. The heuristic of choosing the

most common sense of a term sense is well-known to outperform many of the supervised WSD systems which take surrounding context into account: the high performance of the first sense-based system is due to the skewed frequency distribution of word senses. Nevertheless, because each of the terms that constitute a restaurant NE may have more than one sense, the resolution-based features may be incorrectly constructed due to terms being interpreted in a context other than the correct one. For instance, the resolution-based feature of a restaurant NE [안동(Andong(city)), 국시(noodle or policy)] composed of the ambiguous term 국시(noodle or policy), may be incorrectly constructed as [location, policy] rather than the correct feature [location, food]. Future work should seek to resolve the issue of term ambiguity so as to avoid noisy feature construction.

The last source of NER error was errors stemming from the KLT morphological analysis, which make it difficult for the system to select the correct class. In our experiment, NER errors due to morphological errors accounted for 2.02% of the total errors. When the KLT decomposes a restaurant NE into erroneous terms, the decomposed terms tend to be unknown to the training data and the semantic tree, which contributes to recognition errors. Therefore, the number of KLT errors will influence the precision of restaurant NER. From this error analysis, we could see that 9.12% of the errors (insufficient data) can be easily fixed by human editors as long as there is a human operator to fix these errors continuously, so that the performance of this system can be highly improved and it can be deployed in a practical web site.

5. Conclusion

The main thrust of the present study was to develop a novel method to classify restaurant NEs based on a multi-resolution approach. The clearest trend observed in the results of our experiments was the striking effect of resolution-based feature construction. The results obtained using the proposed method indicated that the concepts in the final resolution could be used as the key features to improve the overall NER performance.

References

[1] T.H. Kim and others, "Proper noun extraction using data sets," *Proc. of the 12th Annual Conference on Human and Cognitive Language Technology*, pp.11-18, 2000.

[2] T.G. Noh and S.J. Lee, "Extraction and classification of proper nouns by rule-based machine learning," *Proc. of Korean Information Science Society*, pp.170-172, 2000.

[3] K.H. Lee, Study on named entity recognition in Korean text, *Master's thesis, KAIST*, Korea, 2000.

[4] K.J. Lee, D.G. Lee, H.C. Rim, S.J. Lim, "Fine grained classification of named entities using machine learning and dictionary," *Proc. of 30th Korean Information Science Society*, 2003.

[5] C.G. Lee, Y.G. Hwang, H.J. Oh, S. Lim, J. Heo, C.H. Lee, H.J. Kim, J.H. Wang, M.G. Jang, "Fine-grained named entity recognition using Conditional Random Fields(CRF) for question answering," *Proc. of 18th Annual Conference on Human and Cognitive Language Technology*, pp. 268-272, 2006.

[6] D. Nadeau and S. Sekine, "A survery of named entity recognition and classification," *Linguisticae Investigationes*, vol.30, no.1, pp.3-26, January 2007.

[7] Restaurant Row, at http://www.restaurantrow.com/.

[8] Menupan.Com, at http://www.menupan.co.kr.

[9] C.G. Kim, "Design and Implementation of the Compound Noun Segmentation Algorithm Based on Statistical Information," *Int. Journal of Fuzzy Logic and Intelligent Systems*, vol.4, no.3, pp.306-310, 2004.

[10] KLT, at http://nlp.kookmin.ac.kr/HAM/kor/.

[11] C. Fellbaum and others, *WordNet: An electronic lexical database*, The MIT press, 1998.

[12] WIRE, at http://www.cwr.cl/projects/WIRE/.

[13] JoyFood.Com, at http://www.joyfood.com.

[14] HTML-parser, at http://htmlparser.sourceforge.net.

[15] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, 2002.

[16] G.M. Weiss, "Mining with rarity: a unifying framework," *The SIGKD Explorations Special Issue on Learning from Imbalanced Datasets*, Vol.6, No.1, pp.1-6, 2004.

[17] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," *Proc. of COLING-2002*, pp.390-396, 2002.

[18] A.S. Yoon, S.H. Hwang, E.R. Lee, and H.C. Kwon, "Construction of Korean WordNet, KorLex 1.5," *Journal of KIISE : Software and Applications*, vol.36, no.1, pp.92-108, 2009.

Bo-Yeong Kang
Professor of Kyungpook National University
Research Area: Machine Intelligence Technology, Robotics
E-mail : kby09@knu.ac.kr

Dae-Won Kim
Professor of Chung-Ang University
Research Area: Data Mining, Machine Learning
E-mail : dwkim@cau.ac.kr