A Korean Noun Semantic Hierarchy (Wordnet) Construction

Juho Lee, Koaunghi Un, Hee-Sook Bae and Key-Sun Choi
KORTERM, Dept. of EECS, KAIST
373-1 Guseong-dong, Yuseong-gu
Taejon 305-701, Korea
{mywork, koaunghi, elle, kschoi}@world.kaist.ac.kr

Abstract

Since thesaurus is used as a knowledge resource in many natural language processing systems, it is very useful and necessary for the high quality systems, especially for dealing with semantics. In this paper, we introduce a semi-automatic method for the construction of Korean noun semantic hierarchy by utilizing a monolingual MRD and an existing thesaurus.

1 Introduction

Thesaurus[1] or wordnet takes too much time and effort to construct them manually. In this paper, we introduce a semi-automatic method for the construction of Korean noun semantic hierarchy with lexical mapping to each noun's sense. In this method, an MRD (Hangeul Society, *ed.* 1997) and an existing translated thesaurus (Ikehara, *et al.* 1997) are used. By assigning the semantic category of the existing thesaurus to each sense[2] of the nouns in MRD, we combine these two resources and produce an expanded lexicalized noun semantic hierarchy. The semantic category is assigned first and manual correction is performed in post-processing, so semantic hierarchy is constructed with relatively high accuracy and small effort.

2 Related Works

There were researches on construction of a WordNet using existing WordNet and MRD. In EuroWordNet project, a multilingual database for several European languages was built based on English WordNet. For example, Spanish WordNet was constructed using English WordNet (Atserias, *et al.* 1997, Farreres, *et al.* 1998). They used a Spanish monolingual dictionary and bilingual dictionaries for the construction. They supposed three methods and combined those methods in order to get high coverage. For Korean, Moon (1996) used hypernym information of a Korean dictionary and combined it with Korean translation of the English WordNet. The manual pruning was done during the construction. There was another research for the construction of Korean WordNet based on the English WordNet (Lee *et al.* 2000). They used a bilingual dictionary to link the senses of Korean nouns to the synsets of English WordNet. They applied six heuristics to word sense disambiguation and combined each heuristic with decision tree.

This work is different from preceding works in three points; 1) our Korean noun semantic hierarchy is a lexical map including hierarchy and other relationships, not a WordNet; 2) since it is based on a Japanese thesaurus and Korean MRD, we have an advantage of similarity between two languages; 3) we link lexical units, elements of thesaurus, with each sense of nouns in MRD.

---

[1] "Thesauri are based on concepts and they show relationships among terms. Relationships commonly expressed in a thesaurus include hierarchy, equivalence (synonymy), and association or relatedness." [...] "WordNet structures concepts and terms not as hierarchies but as a network or a web" (Hodge 2000)

[2] Because the terms, "sense" and "signification" are differently treated by linguists, we clarify that the "sense" indicates the diverse realizations of "signification" corresponding to a lexical unit in this study.

3 Main Method

In order to select nouns to be included in a new semantic hierarchy, we extracted nouns from large-scale corpus. The overall structure is showed in Figure 1. The main method consists of three stages. In the first stage, semantic categories are assigned to the highly frequent nouns using the existing thesaurus. We call this Japanese existing thesaurus as "NTT thesaurus" in this paper. In the second and third stage, the

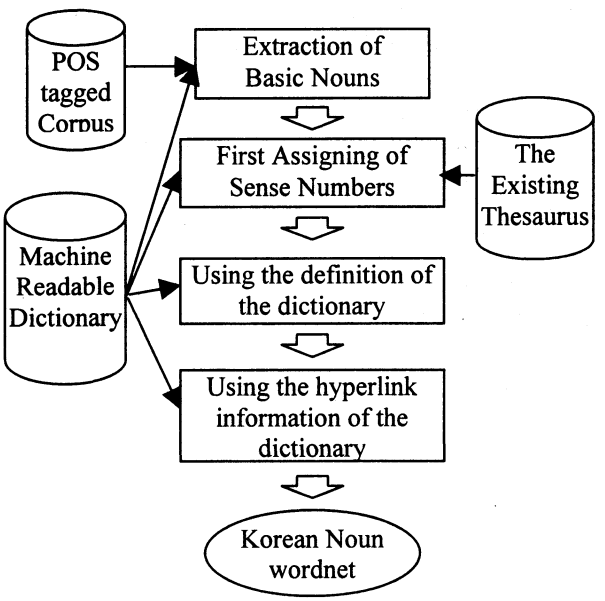

Figure 1. The overall structure

descriptive statements of each sense in the dictionary and hyperlink information of the MRD are used to extend the semantic hierarchy. Then, we will explain the details of each stage.

3.1 Selection of the Nouns

It is difficult to deal with all nouns in MRD[3]. This is why we selected the highly frequent nouns, which will be linked to each semantic category of the thesaurus, from corpus. It is preferable that the semantic hierarchy consists of the nouns, which appear frequently in corpus, for the purpose of high applicability of the semantic hierarchy. In this paper, we selected highly frequent nouns from KAIST POS-tagged corpus[4]. And then, information for highly frequent nouns was extracted from a Korean dictionary (Hangeul society, *ed.* 1997). The highly frequent nouns covered about 91.1% of all nouns in that corpus. The statistics of the highly frequent nouns are as follows: 25,368 unique nouns whose total number of senses are 69,242; the average number of senses per noun is 2.73. NTT thesaurus contains 2,710 hierarchical semantic. Our major task is to assign one of the 2,710 semantic categories to each of the 69,242 senses.

3.2 Assignment of Semantic Categories

In this stage, semantic categories are assigned to each sense of the highly frequent nouns with reference to the noun list of NTT thesaurus. First of all, we translated the noun list of NTT thesaurus into Korean using Japanese-Korean MT system. And then, experts correct the result of automatic translation. Finally, we arbitrate manually the cases of abnormal assignment between the languages. In spite of this process, the translation makes many problems; the most difficult problem is due to the difference of concept division system. For example, in Japanese thesaurus, words concerning "going" or "sorting" have more

[3] *Urimal* Korean dictionary has about 300,000 nouns.

[4] It is composed of 10 million eojeols.

branches than in Korean language, and vice versa for word *root*. In addition, in the course of adapting a Japanese thesaurus to Korean language, we also find many problems. In NTT semantic category hierarchy, the number of word *furniture* is <895>. This word contains its hyponyms <*desk* 896>, < *chair* 897>... <*fireplace* 900>. For the Japanese, the word *fireplace* is understood as a sort of furniture, while the Korean treat this word as a part of the kitchen. These problems issue from the difference of thinking and culture.

Then we can assign semantic categories by matching the highly frequent nouns to translated noun list of NTT thesaurus. If we don't find a noun in the NTT noun list, we follow hypernym of the highly frequent noun and use that hypernym until the matching succeeds. Hypernym of the noun can be extracted automatically from the descriptive statements of the monolingual dictionary. Hypernym is considered as the head noun of the first noun phrase in the descriptive statement for the definition of the noun. In Korean, the head noun is located at the last position of noun phrase generally, so it is easy to find the head noun without parsing. We used following rules to find hypernym.

- The last noun of first noun phrase in the definition is hypernym
- We used 27 lexical patterns. For example, if the pattern such as '**A** *eui hana* (a kind of **A**)' or '**A** *eui ilbu* (a part of **A**)' is applied, **A** is hypernym.
- If the sense of extracted hypernym is too general, that hypernym is discarded.

However, this method ignores the sense of the noun itself, the candidates of semantic categories are not assigned to the sense but to the noun. Moreover, there can be some errors in Korean translation of Japanese thesaurus. In post-processing, word sense disambiguation was done manually to assign proper semantic categories to each sense of the noun and the translation errors were also removed. The principle of word sense disambiguation is as follows.

- Assigning the sense numbers based on the definition of the dictionary.
- It is possible to assign many sense numbers to one sense.
- If the two sense numbers is assigned the same sense and one sense number is a descendent of the other, we choose only a descendent

Two people performed independently the same post-processing. The results of them were compared to each other and the only identical part of them was selected for the final semantic category to achieve the high accuracy. A third party examined the different parts of the results and chose the proper ones. We assigned semantic categories to 29,637 senses for 19,663 nouns in this stage. However, there was a little lower applicability than what we expected because of the translation errors and the discrepancy between the highly frequent nouns and the noun list of NTT thesaurus. The hypernym information was not enough to compensate these defects. It makes us use other information additionally in the next stage for the higher applicability.

3.3 Use of the Definition in MRD

In this stage, we use the result of the previous stage and the definition of the dictionary to expand the preliminary semantic hierarchy in preceding section.

3.3.1 Approach

We used an information retrieval technique on the assumption that the senses, which are in the same semantic category, are defined by similar words in the dictionary (Chen 1998). For the senses to which we had assigned semantic categories in the previous stage, we clustered the definitions of the senses into semantic categories. A cluster of the definitions per semantic category was made. Each cluster corresponds to the document of IR and the definition of the sense corresponds to the query of IR. Assigning proper semantic categories to the sense can be viewed as retrieving relevant documents for the query. We have already assigned semantic categories to the part of senses in the previous stage so we can assign semantic categories to the rest of the nouns by this approach based on the previous results.

3.3.2 Algorithm

We must compute the similarity between the definition and the cluster to retrieve relevant clusters for that definition. We used simply *tf* (term frequency) and *idf* (inverted document frequency) to compute the similarity and gave an extra weight to hypernyms. The similarity between the definition Q and the cluster C_i was computed by this equation.

$$\sum_{t_j \in Q} g(t_j) \times tf_{ij} \times \log\left(\frac{N}{df_i}\right)$$

t_j: Content words which is in the definition Q
$g(t_j)$: function to give a weight for hypernym

$$g(t_j) = \begin{cases} w & \text{if } t_j \text{ is hypernym } (w=2 \text{ is used}) \\ 0 & \text{otherwise} \end{cases}$$

tf_{ij}: The frequency of word t_j in cluster C_i
N: The number of clusters
df_i: The number of clusters where word t_j appears

We computed the similarity between the sense and each cluster and found most relevant cluster. And now, we could find proper semantic categories of that cluster.

3.3.3 Experiment

We clustered the definitions, to which the semantic category had been assigned in the previous stage, by the semantic category. We chose 150 senses, which are not yet involved in the assignment of semantic category, randomly as experiment data. Figure 2 shows the number of senses of which the relevant cluster appeared in higher n^{th} rank of the result. As n increases, the number of senses converges. When we gave 10 candidates per sense, the recall was 61.3%. This result shows that this second stage is useful for extending the semantic hierarchy. Post-processing is also required in this stage by the same method in 3.2. In this stage, we added 19,263 senses for 9,006 nouns to the semantic hierarchy. The Average sense number per noun, which was assigned in this stage, is bigger than that of the first stage. In other words, the method of the second stage is more useful in dealing with polysemy.

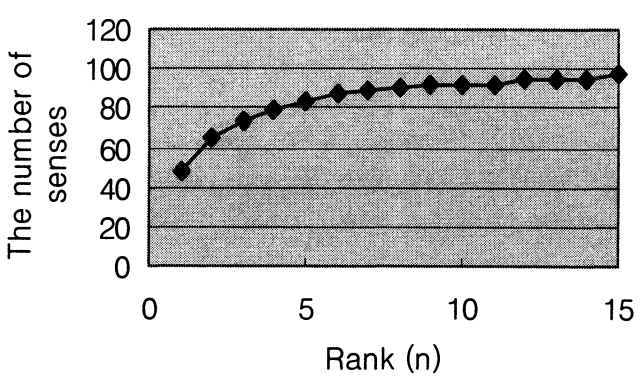

Figure 2. The number of senses for the rank

3.4 Using the Hyperlink Information

In this stage, we use the hyperlink information to extend the semantic hierarchy. Our structured version of Korean dictionary has hyperlink information such as synonym, abbreviation, antonym, *etc.* It is reasonable that the two senses, which are linked by this hyperlink information except antonym, belong to the same semantic category. So we expand the semantic hierarchy using this properties based on the

previously accumulated results and hyperlink. The post-processing is not necessary in this stage. In this stage, we added 7,623 senses for 4,658 nouns to the semantic hierarchy.

4 The Integrated Browsing Tool

The manual post-processing is the important part of this construction. We made an integrated browsing tool for the dictionary and the semantic hierarchy.
Figure 3 shows this browsing tool. This tool is constructed with WWW interface and divides into four frames for input, semantic hierarchy browsing, MRD browsing and semantic category browsing.

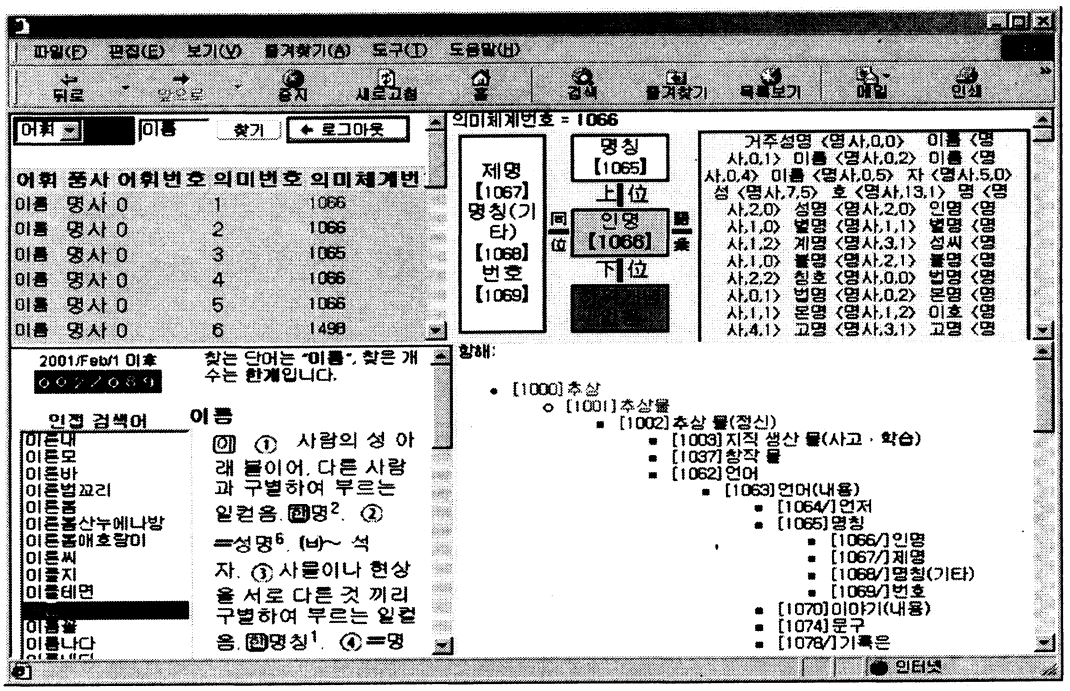

Figure 3. An integrated browsing tool

In the upper-left frame, we input the queries. We can search the result by word, semantic category number and semantic category name. The result of the search is shown as the table form in this frame.
In the upper-right frame, we can browse the semantic hierarchy. This frame consists of five boxes. The center box means the current semantic category. The upper box means the hypernym and the lower box means the hyponym. The left box means the sibling and the right box means the sense of the dictionary that belongs to the current semantic category. We can browse the thesaurus by clicking each element of these five boxes.
The lower-left frame shows the content of MRD. We can confirm the definition of the dictionary easily.
The lower-right frame shows the hierarchical structure of the thesaurus. We can show or hide the subordinate semantic category by clicking the parent semantic category.
These four parts doesn't work independently. They operate together. The change of one frame affects every other frame. This browsing tool was much helpful to the post-processing during the construction of Korean noun semantic hierarchy.

5 Conclusion

In this paper, we introduced a semi-automatic method for the construction of Korean noun semantic hierarchy. This method uses a MRD and an existing thesaurus. The method consists of three stages. In the first stage, semantic categories are assigned to the highly frequent nouns using an existing thesaurus. In the second and third stage, the definitions of the dictionary and hyperlink information are used to

expand the semantic hierarchy. We constructed Korean noun semantic hierarchy, which has 56,523 senses for 23,823 nouns and 2,710 hierarchical semantic categories.

References

Atserias, Jordi, et al. 1997. Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In *Proc. of International Conference on Recent Advances in NLP.*

Chen, Jen Nan and Jason S. Chang. 1998. Topical Clustering of MRD Senses Based on Information Retrieval Techniques. *Computational Linguistics Volume 24, Number 1.*

EuroWordNet home page. http://www.hum.uva.nl/~ewn

Farreres, Xavier, et al. 1998. Using WordNet for building WordNets. In *Proc. of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems.*

Hangeul Society, *ed.* 1997. *Urimal Korean Unabridged Dictionary,* Eomungag.

Hodge, Gail M. 2000. *Systems of knowledge organization for digital libraries : beyond traditional authority files.* Washington, DC : Digital Library Federation, Council on Library and Information Resources.

Ikehara Satoru, *et al.* 1997. *The Semantic System*, volume 1 of *Goi-Taikei - A Japanese Lexicon.* Iwanami Shoten.

Lee, Changki and Geunbae Lee. 1999. Using WordNet for the Automatic Construction of Korean Thesaurus. In *Proc. of 11th Hangeul & Korean Information Processing*, pp. 156-163 (in Korean).

Lee, Changki, *et al.* 2000. Automatic WordNet mapping using word sense disambiguation. In *Proc. of Joint SIGDAT Conference on EMNLP/VLC.*

Miller, George A. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography.*

Moon, Yoo-jin. 1996. Design and Implementation of WordNet for Korean Nouns. Journal of KISS (c) 2/4, pp. 437-445 (in Korean).