Korean-Chinese-Japanese Multilingual Wordnet with Shared Semantic Hierarchy

Key-Sun Choi, Hee-Sook Bae, Wonseok Kang, Juho Lee,
Eunhe Kim, Hekyeong Kim, Donghee Kim, Youngbin Song[1], Hyosik Shin

KAIST, Korterm, Bank of Language Resources
Computer Science Division, 373-1 Guseong-dong Yuseong-gu Daejeon 305-701 Korea
kschoi@cs.kaist.ac.kr

Abstract

A Chinese-Japanese-Korean wordnet is introduced. It is constructed based on a shared semantic hierarchy that is originated from NTT Goidaikei (Lexical Hierarchical System). Korean wordnet was constructed through the semantic category assignment to every sense of Korean words in a dictionary. Verbs and adjectives' senses are assigned to the same semantic hierarchy as that of nouns. Each sense of verbs is investigated from corpora for their usage, and compared with Japanese translation. Chinese wordnet with the same semantic hierarchy was built up based on the comparison with Korean wordnet. Each sense of Chinese verb corresponds to Korean with its argument structure. These works have lasted since 1994.

1. Introduction

A Chinese-Japanese-Korean wordnet has been built on one common semantic hierarchy. Here, we will use "wordnet" in a general sense of the network of words, not the sense of WordNet (Miller, 1995). This semantic hierarchy is originated from NTT Goidaikei (Ikehara, et al., 1997) which consists of 2,710 hierarchical semantic categories. We will use the term "concept" for "semantic category" in this paper. The number of concepts in CoreNet is more specified into 2,954. The reason for increased concepts is as follows: first, it is to reflect the concepts that were found in Korean. Second, CoreNet uses the same semantic hierarchy for nouns and predicates, although NTT Goidaikei uses different concept system for nouns and predicates.

The use of the same semantic hierarchy for nouns and predicates has several advantages. First, the surface forms of nouns and predicates share the similar one, especially in Chinese words. In case of Korean and Japanese, the typical formation is like "do+N" in English like "N+suru" in Japanese and "N-hada" in Korean. Second, the language generation from conceptual structures takes freedom to choose the surface form whether it chooses noun phrases or verb phrases.

This computational works experienced heuristics and trial-and-errors as well as semi-automatic approaches. A brief introduction to the principles in CoreNet is described in section 2. Its construction procedures will be shown in section 3. Section 4 is reserved for explaining consideration points. CoreNet has been built up based on several linguistic resources. Among them, Korean word senses are mainly based on (Choi, 2000) and (Hangeul Society, 1997). The set of Chinese vocabulary mainly depends on (Yu, 1999).

2. Principles

CoreNet has been constructed by the following principles: (1) word sense mapping to concept, (2) corpus-based, (3) multi-lingualism, (4) mono-concept system for multi-languages

2.1 Sense Mapping to Concept

The major purpose of CoreNet is to resolve semantic ambiguities by two functionalities. Every sense of words in the dictionary (Hangeul Society, 1997) is mapped to at least one concept. For example, each sense of word "*school*" is mapped into three concepts under PLACE, ORGANIZATION, and BUILDING. The other functionality is to give the syntactic-semantic structure for predicates, which is based on the predicate-argument structure.

For example, a Korean verb "*gada*" has 17 senses in the dictionary (Hangeul Society, 1997) which are mapped into concepts GOING, LEARNING, SERVICE, DELIVERY, PROGRESS, CONTINUATION, ENTHUSIASM, SWEEP, and so on. This set of concepts for predicates shares the same with those of nouns.

On the other hand, each predicate has its own argument structure. For example, "*gada*" has argument structures with concepts and Japanese translation is as follows:

- going([human,mammal,vehicle]=subj), "*iku*"
- learning([human]=subj,[teacher]=dat), "*iku*"
- delivery([information]=subj,[human]=dat), "*tutawaru*"
- progress([time]=subj), "*sugiru*"
- continuation([relation]=subj,[year]=obj), "*tuduku*"
- enthusiasm([gaze]=subj,[girl]=dat), "*iku*"
- sweep([emotion]=subj), "*kieru*"

2.2 Corpus-based Usage

The set of vocabulary and their senses are extracted from KAIST corpus (Choi, 2000). For example, all argument structure of "*gada*" in the former section is extracted from the corpus as follows:

- GOING([*HORSE*/MAMMAL, *BUS*/VEHICLE]=SUBJ)

HORSE and *BUS* are extracted terms from corpus and MAMMAL and VEHICLE are concept names mapped from words *horse* and *bus*. This causes the more specified sense categorization than those of dictionaries.

2.3 Multi-Lingualism

[1] Song, Youngbin's current address is Ewha Womans University.

All of concepts are aligned among three languages: Japanese, Korean and Chinese. All of words in noun and predicate of three languages are categorized into one common concept hierarchy. Verbs of three languages are also linked each other based on senses and concepts. A partial list of concepts for Chinese verb "去" [qù] is as follows: (*italicized* words are Korean translation.)

- GOING - *gada*
- DELIVERY – *bonaeda*
- EXCLUSION – *eobsaeda*

Seven Korean sennses for Chinese verb "gǎo" is listed as follows[2]:

搞 [gǎo] [V]
1. 마련하다 (prepare, reserve)
[N1] 他(23,48) [V] [N3] 票(932)|电视机(970)
2. 만들다 (make)
[N1] 他(23,48) [V] [Aux]了 [N3] 方案(1036,1436)
3. 하다 (do)
[N1] 他(23,48) [V] [N3] 设计|施工| 生产|工作
4. 경영하다 (manage)
[N1] 他(23,48) [V] [N3] 工厂(439)
5. 맡다 (take in charge of)
[N1] 他(23,48) [V] [N3] 总务(326)
6. 찾다 (search)
[N1] 他(23,48) [V] [N3] 对象(74)
7. 맺다 (form, contract)
[N1] 他们(25,2606) [V] [N3] 关系(1684,2444)

2.4 One Concept System

In general, concept systems and word nets are constructed for words in noun. However, CoreNet shares one concept systems for nouns, verbs, and adjectives. Furthermore, one concept systems have been used and updated to keep three languages share one.

3. Automatic Procedure

An initial procedure of CoreNet construction is summarized from Lee, et al. (2002). Then the manual procedures will be described. See figure 1.

3.1 Overall Approach

A basic set of words are selected from both a frequency-based vocabulary list corpora comparing with the existing basic Korean word sets. About 50,000 general vocabularies are selected for CoreNet word entries.

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

[2] [N1], [V] and [N3] mean the first position (noun), verb itself, and the third position (noun) respectively. The typical usages are shown with extracted from the resources. The number in the parenthesis is the semantic category number.

3.2 Semantic Category Assignment

Semantic categories are assigned to each sense of the basic set of words with reference to the word set that belongs to a semantic category. First of all, we translated all of the Japanese words under NTT Goidaikei into Korean words using a Japanese-Korean electronic dictionary. Experts correct the result of automatic translation. Then we manually correct the erroneous assignments between two languages. In spite of this process, we have many problems.

The most difficult problem is due to the difference of concept division system. For example, in Japanese, words concerning GOING or SORTING have more branches than in Korean language, and vice versa for ROOT. In addition, *FURNITURE* has its hyponyms *DESK*, *CHAIR,* and *FIREPLACE*, while the Korean treats this word as a part of KITCHEN. These problems issue from the difference of thinking and culture.

Then we assign semantic categories by matching the Korean words with the translated word list under the NTT Goidaikei's semantic category. If no translated word is found in the translated word list, we follow the genus term of the word that is extracted from the descriptive statements of a machine-readable dictionary. 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 word and the translation errors were also removed. 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.

3.3 Dictionary Use

In this stage, we use the hyperlink information. 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 concept.

4. Considerations

This section describes what we had to consider and decide.

4.1 Underspecified Sense and Multiple Concept Mapping

A word is assigned to the multiple numbers of concepts for each sense of the word. For example, *school* is an "institution for the instruction of students." This word *school* is mapped to three concepts for LOCATION, ORGANIZATION, and FACILITY as shown before. However, if the sense in the mother dictionary is underspecified, one sense of word must be mapped to the multiple numbers of concepts. The word *school* is one of underspecified examples of senses.

4.2 Verbal Noun

A word in verb is assigned to concepts after it is transferred to its noun form. For example, "*write*" is transformed to its noun form "*writing*" that is mapped to a

concept WRITING under EVENT. An adjective "*be wise*" is transformed to "*wisdom*" that is mapped to PROPERTY under CAPABILITY that is again under ATTRIBUTE. Consider an adjective "*be wide*". A sense is mapped to POSITIVE PERSONALITY, EXTENT/ LIMITS, and WIDTH (under the concept UNIT OF QUANTITY), respectively.

4.3 Concept Splitting

Whenever we found the inconsistency among nodes of concepts, a node may be added. For example, BODY had three subconcepts in NTT concepts: ARM, LEG, and HEAD. But, a word "*back*" cannot be assigned to any subconcepts. At least, OTHERBODY should be added to the fourth subconcepts under BODY.

5. Word Distribution in Concepts

5.1 Noun Distribution

The topmost concept labels are CONCRETE and ABSTRACT. A little less than half entries (20626/51614) are located in CONCRETE, that is, 20626 entries among 51614 senses. Let 'big concept' be the concept which has entries more than others. Such concepts are PLANTS, FLOWER/WILD GRASS, SUBSTANCE (PART), WATER, PERSONAL EFFECTS, CLOTHES (MAIN), VEGETABLE, PUBLICATION, WEAPON, SCIENCE AREA/DEPARTMENT, PERSON'S NAME, WORD, DOCUMENTS, MIND, and UNIT.

5.2 Verb Distribution

Because verbs are mapped into the same concept system as the nouns, all the verbs are under EVENT. The following three concepts are mainly shown in this order: HUMAN ACTIVITY, FACT/PHENOMENA, and NATURAL PHENOMENA.

5.3 Adjective Distribution

Adjectives are also under EVENT. But they are under the different concepts like RELATION, PROPERTY, STATE, and APPEARANCE.

6. Differences

Some senses of words in one language cannot be found in the other language. In case of Chinese-Korean translation, we could find that the following Chinese words have no translation in Korean: "手感(shǒugǎn)": feeling by hand, and "省优(shěngyou)": quality awards from provinces (authorities)

7. Examples

Figure 2 is a browsing window. In the left-up side, 'lexical map sorted by entry' consists of five tuples: entry, part-of-speech, homonym nuber, sense number, and concept number. The homonym number is the serial number of one entry in the dictionary, while the sense number is the serial number in one homonym. Their dictionary plane is shown in the left-bottom side, and their concept plane is in the right-up side, where all of the links are drawn. In the right-up side, 'lexical map by concept hierarchy' shows the vocabulary lists under the current semantic category. (Here, the terms "semantic category" and "concept" are used for the same purpose.)

Figure 3 shows a screenshot of Korean-Chinese noun wordnet. The screen has four windows. Left-Up window shows the correspondence between Japanese and Korean words and concept numbers. The left-down window contains the word senses and definition in the dictionary (Hangeul Society, 1997). Right-Up window is to show all words under a concept QUANTITY numbered 2588. The right-down window is one partial list of concept hierarchy.

8. Conclusion

CoreNet has been constructed while its necessary corpora and lexical database are also developed. The starting point is to use the skeleton of NTT Goidaikei (Ikehara, et al., 1997). Then Korean version of noun systems was developed. Although NTT Goidaikei used the different semantic categories for predicates, CoreNet tried to use the same one. Another different thing between CoreNet and NTT Goidaikei is that CoreNet has a mapping between word senses and concepts. Multilinguality is also embodied to see one conceptual system that can be used for different languages.

Acknowledgements

This work was supported by the Ministry of Science & Technology in Korea under the project "Korea Information Base Systems" (1995-2000), and managed by BK21 fund of Ministry of Education (2001-2003).

References

Ikehara, S. *et al*. (1997). The Semantic System, volume 1 of Goidaikei – A Japanese Lexicon. Tokyo: Iwanami Shoten.

Choi, K.-S. (2000). KAIST KIBS Corpus. http://morph.kaist.ac.kr/kcp/.

Hangeul Society, *ed*. (1997). Urimal Korean Unabridged Dictionary. Seoul: Eomungag.

Lee, J.H., Un, K. & Choi, K.-S. (2002). =Semi-Automatic Construction of Korean Noun Thesaurus by Utilizing Monolingual MRD and an Existing Thesaurus. In Proceedings of the 16th Pacific Asia Conference (pp. 47-50). Jeju.

Yu, S. (1999) Modern Chinese Grammar Information Dictionary. Peking Univeristy Press.

Miller, G. (1995) WordNet: a Lexical Database. Communications of the ACM, 38(11), 39-41.

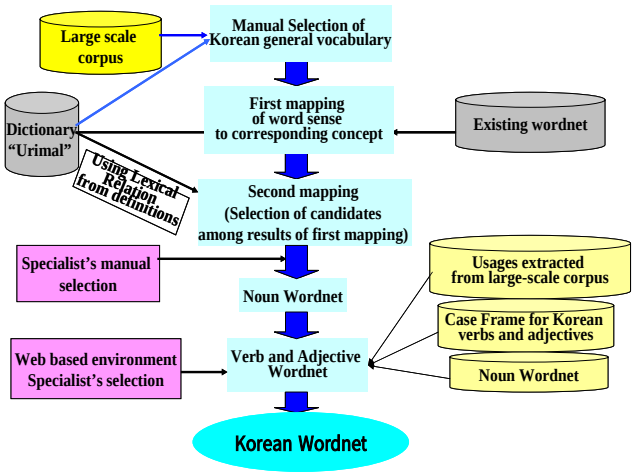

Figure 1: Concept Assignment of Word Senses

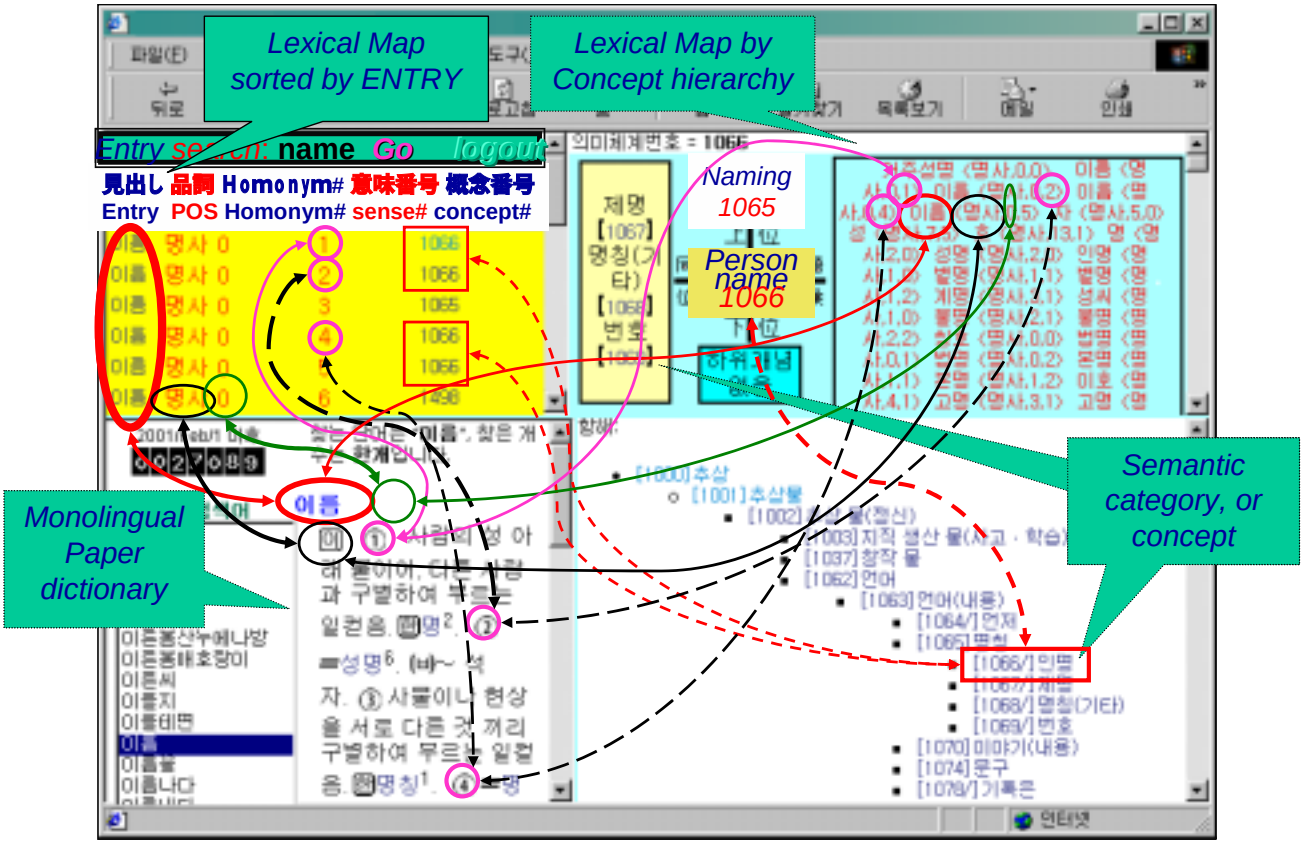

Figure 2: Browsing Window

Figure 3: Korean-Chinese Wordnet