Project Report on a Korean Science & Technology Thesaurus with Conceptual/Relational Facets

Hanmin Jung and **Won-Kyung Sung** and **Dong-In Park**
Division of Information System, Korea Institute of Science and Technology
52 Eueon-dong, Yuseong-gu, Taejeon, KOREA 305-806
{jhm, wksung, dipark}@kisti.re.kr

Abstract

Our project has a long-term plan to construct a Korean science & technology thesaurus from 2005 to 2010. For designing an elaborated thesaurus, we introduce conceptual and relational facets which are excluded or partially included in WordNet, Core-Net, and other thesauri constructed by Chung *et al*. (2002) and Lee *et al*. (2000).

About Our Thesaurus

Cont is represented as a set of a descriptor (USE) and non-descriptor(s) (UF) like a synset on WordNet. For example, ‘NOR flash memory’ concept has “NOR형 플래시 메모리 (*NOR flash memory*)” as a descriptor, and “노어 플래시 메모리 (*NOR flash memory*)” and “코드 플래시 메모리 (*Code flash memory*)” as non-descriptors. Descriptor is an indispensable term with one or more conceptual facets (CF). Non-descriptor is omissible in the case that a descriptor has no synonym. Conceptual facets, representative attribute categories of concepts, are attached to each concept. It implicitly means conceptual facets of UFs should be the same as that of corresponding descriptor. It is a restriction to build our thesaurus. BT is a broader concept and NT is a narrower concept connected with it using relational facets. The facets are the viewpoints that a broader concept looks at its narrower concept. We think the major reason that previous thesauri in Korea have been criticized by users is the lack of receptive capacity for numerous users’ viewpoints. In our thesaurus, three kinds of relational facets lie on the edge of a BT-NT relation. An NT can have several BTs in the case of multiple inheritances. Category relational facet (CRF)[1] and attribute relational facet (ARF)[2] are obligatory, but thematic-role relational facet (TRF)[3] is applied only when the head of NT is a predicative noun. Attribute keyword (AK) is defined as an additional feature of NT which discriminates from BT, for example, if BT includes “cellular phone” and NT “camera phone,” then the attribute keyword of NT would be “camera.” It is the criterion to determine attribute relational facet and thematic-role relational facet on a BT-NT relation. The keyword is also helpful to determine synonym candidates whether they are in a synonym set or not. The following figure shows a thesaurus example with four concepts and their conceptual and relational facets.

Thesaurus construction process includes automatic term extraction, term selection, sense & domain determination, conceptual facet attachment, concept information filling, BT-NT relation establishment, and relational facet attachment. We tried to satisfy Nilson’s the requirement defined by Nilsson *et al*. (2000) for building semantic metadata; subjective, evolving, extensible, flexible, conceptual, and distributed.

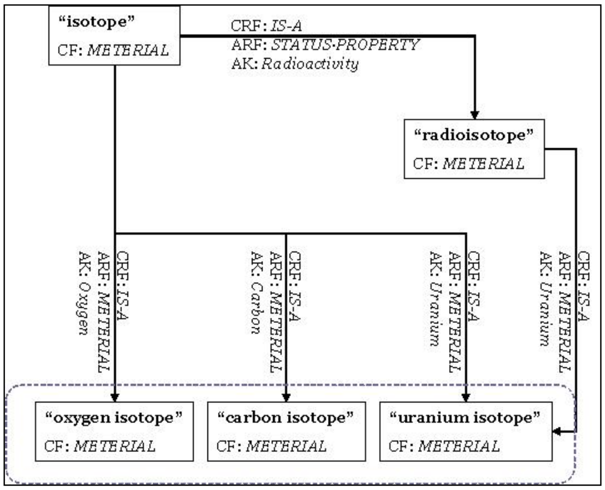
As the first step to build the thesaurus, we extract science & technology terms from general corpora including newspapers, magazines, and newsletters with period information. Term dominance value (TDV) is defined by Jung *et al*. (2005) to rank the terms by means of term dominance trend that implies what terms are increasingly used in recent years. We also define the life cycles

[1] It represents the type of a BT-NT relation. The facet corresponds with NT, NTi, and NTp defined in other thesauri.

[2] It means the representative category for an attribute keyword on a BT-NT relation. Most of the members of attribute relational facets are shared with those of conceptual facet.

[3] It indicates the semantic relationship between a predicative noun (head), which is equivalent to a predicate in Korean, in NT and an attribute keyword on a BT-NT relation.

of terms and their stages. Each term is assigned to one of new, growing, steady, declining, disappeared, and recycled stages. The experimental results of Jung *et al*. (2005) showed that dominant terms[4] have higher survival probability than declining and disappeared terms. Thus about 50,000 dominant terms were selected as seeds to build the thesaurus.



> We also have another principles including the followings to select science & technology terms; English-Korean mixed terms and transliterators are permissible (e.g. "DVD 플레이어 (*DVD player*)" and "더블데이터레이트 (*DDR*)"). Dictionary terms are permissible if their domains are clear (e.g. "프린터 (*Printer*)" in computer science). Terms of which domains are easily identified are permissible (e.g. "공인인증 (*Public authentication*)" in computer science). Terms including normal predicative nouns are not permissible (e.g. "S/W 개발 (*S/W development*)").

Sense & domain determination step chooses the domains of terms using statistical information from 13 science & technology domains. We gathered domain corpora and extract answer terms. In the case of matching with these answers, the above manually selected terms become have their senses and domains.

A concept consists of scope notes, English term(s), and usage example(s) as well as conceptual facet(s), a descriptor, and non-descriptor(s). Since it can be relatively de-fined by its information and various relational facets, describing scope note is restricted within a communication method between human thesaurus constructors for harmonious collaboration.

Relation establishment is to connect a broader concept with a narrower concept. We use morphological analysis and keyword-in-context (KWIC) indexing for a rough term clustering. A lot of combination patterns for the constituent elements of com-pound words play a major role on deciding BT-NT relations.

Related concept (RT) which is usually introduced in other thesauri is excluded in ours because Korean experts do not agree its usefulness for information retrieval and inference. The absence of strict guidelines for applying related concepts would be a major reason in Korea. We have a plan to prudentially introduce RT relations defined by clear relation types.

We currently define about 16 members for attribute relational facet including *FIELDTHEORYMETHOD*, *METERIAL*, and *LOCATIONSPACE*. Conceptual facet consists of 15 members except for *INSTANCE* from the attribute relational facets[5]. Category relational facet includes *IS-A*, *HAS-PART*, and *INSTANCE*. Thematic-role relational facet, originated from Fillmore's case frames, consists of 9 members including *SOURCE*, *OBJECT*, and *INSTRUMENT*. Attaching them to BT-NT relations and concepts are the most difficult task of this project. We expect that variations for applying the facets between human constructors would be minimized by thoroughly describing guidelines for all the above steps.

We are also constructing an ontology and RDF data for national R&D base information. For query term expansion during inference, they will be dynamically connected with the thesaurus for users' queries. We hope to contribute the paradigm of thesauri and lexical semantic networks including WordNet by proving the usefulness of the facets as various restriction criteria for the expansion.

References

Chung Y., Kim M., Lee J., Han S., and Yoo J. (2002) *An Integrated Ontological Approach to Effective Information Management in Science*

[4] Terms with positive TDV values, and with on new, growing, steady, recycled stages.

[5] For naming the attribute relational facet members, we tried to keep on about 4.5 for the average depth of corresponding synsets on WordNet.

and Technology, Journal of Korea Society for Information Management, Vol. 19, No. 1.

Jung H., Koo H., Lee B., and Sung W. (2005) *Acquiring Dominant Compound Terms to Build Korean Domain Knowledge Bases*, Proceedings in the 4th Annual International Conference on Computer and Information Science.

Lee C., Lee G., and Seo J. (2000) Automatic *WordNet Mapping Using Word Sense Disambiguation*, Proceedings in Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Nilsson M., Palmér M., and Naeve A. (2000) *Semantic Web Metadata for e-Learning – Some Architectural Guidelines*, Proceedings in the World Wide Web Conference.