

Enriching Core Ontology with Domain Thesaurus through Concept and Relation Classification

Jin-Xia Huang, Ji-Ae Shin^{*}, Key-Sun Choi

Semantic Web Research Center/Computer Science Department,
Korea Advanced Institute of Science and Technology, Daejeon, Korea

^{*} Information and Communications University, Daejeon, Korea
hgh@world.kaist.ac.kr, jiae@icu.ac.kr, kschoi@cs.kaist.ac.kr

Abstract. This paper proposes a methodology for enriching core ontology with the concepts and relations in the domain thesaurus. The concepts of the domain thesaurus are classified into concepts in the top-level of core ontology, and relations between broader terms (BT) - narrower terms (NT) and related terms (RT) are classified into semantic relations defined on the core ontology. To classify concepts, a frequency-based approach is complemented with a similarity-based approach. To classify relations, two techniques are applied: (i) for the case of insufficient training data, a rule-based module is for identifying *is-a* relation out of *not is-a* ones; a pattern-based approach is for classifying non-taxonomic semantic relations from *not is-a*. (ii) For the case of sufficient training data, a maximum-entropy model is adopted, where kNN approach is for noisy filtering of training data. A series of experiments show that performance of the proposed systems are quite promising and comparable to judgments by human experts.

1 Introduction

Ontology includes concepts, semantic relations between concepts, instances, and axioms on the domain. Existing lexical knowledge bases, such as thesaurus, contain semantic information on terms/concepts and taxonomic relations. This is the reason why thesaurus is frequently used for ontology construction. However, the hierarchy of the thesaurus contains not only *is-a* relations and *part-whole* relations between BT and NT, but also other related relations between BT and RT [1]. It is more serious in domain thesaurus. For example, in Figure 1, each hierarchy in domain thesaurus, Inspec, includes both *is-a* and other non-taxonomic relations without clear identification. We call this kind of hierarchy BT-NT/RT hierarchy to distinguish it from the taxonomic hierarchy.

To build ontology with a domain thesaurus, the BT-NT/RT hierarchy has to be mapped into the taxonomic hierarchy for ontology, and the BT-NT/RT relations in the thesaurus have to be classified into semantic relations. Let part A in Figure 2 indicate the BT-NT/RT hierarchies in the domain thesaurus, then as shown in part B, the semantic relations for ontology should be elicited from BT-NT/RT relations in the thesaurus (relation classification). And as shown in part C, the concepts of the

thesaurus should be classified into the categories (top-level concepts) in core ontology (concept classification), to map BT-NT/RT hierarchies of the thesaurus into a taxonomic hierarchy for ontology.

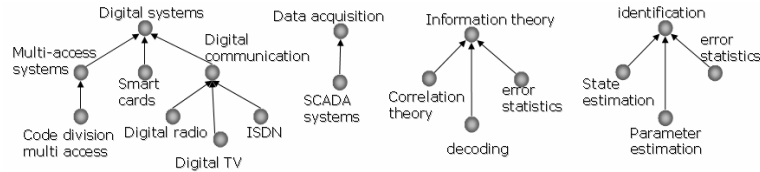


Fig. 1. BT-NT/RT hierarchies in the domain thesaurus Inspec

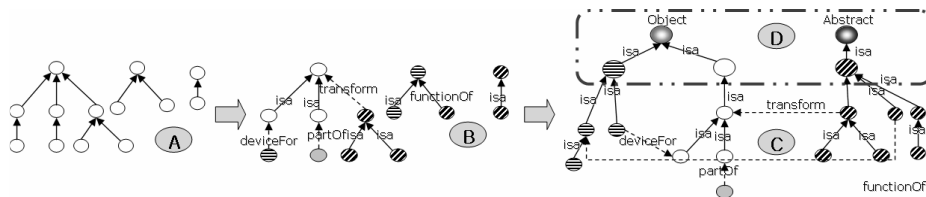


Fig. 2. Enriching ontology with thesaurus by mapping relations (A→B) and concepts (B→C, D).

In this paper, core ontology in the IT domain is given as the target ontology, and a domain thesaurus, Inspec, is used as the source thesaurus. For **concept classification**, the concepts on the top level of the core ontology are considered as semantic categories; the concepts in the Inspec Thesaurus are classified into these target categories. A frequency and similarity-based approach is implemented for concept classification. For **relation classification**, the semantic relations defined in core ontology are considered as relation categories, and the BT-NT/RT relations of the Inspec Thesaurus are classified as specific semantic relations. A set of rules based on lexical information of concepts are proposed to classify BT-NT/RT relations into *is-a* and *not is-a* relations; a pattern-based approach is proposed to classify non-taxonomic semantic relations from *not is-a* relations, in which the BT and NT/RT in relation triples are classified into semantic categories with the approach proposed for concept classification first, then the *not is-a* relation is classified to a specific semantic relation using relation patterns constructed semi-automatically. This is an unsupervised approach, and it is adopted when there is lack of training data. Accompanied with the accumulation of classified semantic relations which can be used as training data,¹ a maximum entropy model (MEM) is proposed for relation classification, where a kNN-based approach is adopted to filter noisy training data.

A series of experiments show that the performances are quite promising, with some of them comparable to agreements among human experts.

¹ This training data is built semi-automatically – the relations are classified with the proposed unsupervised approach first, and then verified by human experts.

2 IT Core Ontology and Domain Thesaurus Inspec

An **IT core ontology** is adopted in this paper as target ontology to be enriched. The core ontology is composed of a domain taxonomy and 185 types of relations defined for the IT domain. The domain taxonomy contains a taxonomic hierarchy with 200 categories as nodes and is part of the general domain thesaurus CoreNet, which contains more than 2,900 categories and 50,000 general vocabularies in Korean [2]. The categories of domain taxonomy are selected mainly according to their popularity score in IT domain. For example, there are many IT terms belong to the CoreNet category *Communication equipment*, therefore its popularity score is high enough to be selected as semantic category for IT domain taxonomy. On the contrary, CoreNet category *Medicines* barely has IT terms belong to it and the popularity score is very low, so it is excluded from IT domain taxonomy (Figure 3).

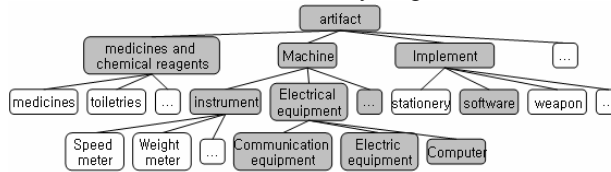


Fig. 3. IT domain taxonomy (the hierarchy of dark nodes) in IT core ontology is part of the general domain thesaurus CoreNet. This figure presents a sub-tree of the *Artifact* category in CoreNet, in which the dark nodes present the categories selected for IT domain taxonomy, and the white nodes are just CoreNet categories.

The semantic relations defined for IT domain ontology have domain and range constraints. In Table 1, e.g. the domain of the relation *theoryAbout* is an instance of category *Theory*, and its range can be an instance of either *Structure* or *Equipment*. The relation triples with domain and range constraints are actually a kind of relation patterns can be used for relation classification. IT core ontology is still under development, and so only part of the semantic relations have domain and range constraints: 258 relation triples are defined for 108 semantic relations (out of 185 types of relations totally) at present.

Table 1. Semantic relations defined for IT domain ontology

Relation	Domain	Range
<i>functionFor</i>	Function	Analysis
<i>functionIn</i>	Function	Logic
<i>functionOf</i>	Function	Plan
<i>theoryAbout</i>	Theory	Structure
<i>theoryAbout</i>	Theory	Equipment
<i>theoryOf</i>	Theory	Information

Inspec Thesaurus. The Inspec Thesaurus [3] is adopted in this paper as the source thesaurus. It covers more than 14 areas including computing, control engineering, electrical and electronic engineering, information technology and physics, etc. Inspec

contains more than 8,300 terms and 15,901 BT-NT/RT relations². The relations are a mixture of BT-NT and BT-RT relations as shown in Figure 1, and there is no labeling (types of relations) assigned to the relations. For example, both (active antennas, antennas) and (antenna theory, antennas) are given as BT-NT relations without labeling.

3 Concept Classification

In this section, we classify Inspec terms into 200 categories of IT domain taxonomy. As the first step, each term is classified into a CoreNet category according to popularity score –a frequency-based approach. Let t be the lexical expression of a term in Inspec and h_t its headword. Assume that h_t has multi-senses m which correspond to CoreNet categories $\{c_1, \dots, c_j, \dots, c_m\}$, respectively. Let w_j represent the popularity score of c_j , and c_t the determined category of t in CoreNet. We have:

$$c_t = c_{h_t} = \arg \max_c \{w_j \mid h_t \in c_j, 1 \leq j \leq m\} \quad (1)$$

The second step is for finding corresponding semantic category of c_t in the IT domain taxonomy C using similarity measurement. Let $C = \{C_1, \dots, C_i, \dots, C_n\}$, where $n=200$ as described in Section 2. The semantic category of t in domain taxonomy, $C(t)$, is classified by Equation (2):

$$C(t) = C(h_t) = \arg \max_C \sum_{i=1}^n Sim(c_t, C_i) \quad (2)$$

Let $depth(c)$ be the depth of a node in CoreNet. The similarity between c_t and category C_i is the maximum reciprocal of the distances between C_i and c_j . The similarity is zero if c_t is not a hyponym category of C_i (Equation 3).

$$Sim(c_t, C_i) = \begin{cases} 0, & \text{if } c_t \text{ is not hyponym category of } C_i \text{ in CoreNet;} \\ \max_{i=1}^n (depth(c_t) - depth(C_i) + 1), & \text{else.} \end{cases} \quad (3)$$

We perform classification on headwords instead of on the terms above. We use the following approach to recognize headword h_t of term t : to a concept in the IT domain, the headword is normally the last word in the compound noun which expresses the concept. If the concept is expressed by only one word, then the word itself is the headword. If the above cases do not apply, we adopt a pattern-based approach to recognize the headword. Let $head(\text{term})$ be a function of headword recognition, thus we derive the following patterns:

- $\langle \text{headword} \rangle \langle \text{prep.} \rangle \langle \text{otherword} \rangle$, where $\langle \text{prep.} \rangle = \{\text{by, in, on, of, from, for, with, about}\}$
 - Ex) $head(\text{learning by example}) = \text{learning}$
- $\langle \text{headword} \rangle _ \langle \text{domain} \rangle$, where $\langle \text{domain} \rangle$ indicates domain information of the concept
 - Ex) $head(\text{network_circuits}) = \text{circuits}$
- $\langle \text{otherword} \rangle _ \langle \text{headword} \rangle$
 - Ex) $head(\text{unsolicited_e-mail}) = \text{mail}$
- $\langle \text{otherword} \& \text{headword} \rangle$, where ‘&’ indicates there is no space between the two words
 - Ex) $head(\text{radiotelephony}) = \text{telephony}$

² Korean lexical expressions and definitions of the Inspec terms are complemented for our work, and used in concept and relation classification.

4 Relation Classification

For relation classification, an unsupervised approach is proposed when there is lack of classified relations that can be used for training; and a supervised approach is introduced with the accumulation of classified relations.

4.1 Rule-based Approach: *Is-a* Relation Classification

Our approach to classify the *is-a* relation relies on two assumptions: **the first assumption** is that two concepts in the BT-NT/RT relation hold an *is-a* relation if they have the same identity; **the second assumption** is that the headword of a lexical expression reflects the identity of its concept. The **identity** of a concept is the essential property that distinguishes the concept from other concepts. For example, “active antenna” has the same identity with its BT “antennas” and so they are in an *is-a* relation. But “antenna testing” has a different identity with its BT’s, “Antennas,” so they are in a *not is-a* relation. Under the two assumptions, a set of rules based on the headwords are proposed to identify *is-a* relations: the rule of the same headword, the rule of transitivity, the rule of tolerance, and the rule of abbreviation.

Rule of the Same Headword. According to the second assumption, if two terms under the BT-NT/RT relation have the same headword, then we know they share the same identity, and so their relation is *is-a* according to the second assumption.

Rule of Transitivity. If the headwords of BT and NT/RT have an *is-a* relation, then according to the above two assumptions and the transitivity of an *is-a* relations, these two terms should also have an *is-a* relation. Suppose we have *isa*(listings, programs) and *isa*(methods, theory) in the IT domain, we have *isa*(JAVA listings, complete computer programs) and *isa*(smoothing methods, filtering theory) as results.

Rule of Tolerance. Some concepts have tolerance in lexical expressions. For example, given the BT “equipment,” it could have many terms like “receivers,” “antennas,” “cameras,” “tubes”, and “transmitters” as its NTs. The terms which have tolerance in lexical expressions include “equipment”, “accessories”, and “applications.”³ If the BT has one of these words as its headword, then it is very likely it has *is-a* relations with its NTs even though these NTs have different headwords.

Rule of Abbreviation. Some concepts have many NTs in abbreviation form. “Languages,” “standards,” and “networks” are examples of such concepts. Similar with the rule of tolerance, if a BT has one of these words as its headword, and its NT is an abbreviation, then most likely they have an *is-a* relation. For example, we have: *isa*(BASIC, high level languages), and *isa*(ISDN, telecommunication networks).

4.2 Pattern-based Approach: Semantic Relation Classification

In Section 4.1, BT-NT/RT relations are classified into *is-a* and *not is-a* relations. In this section, the *not is-a* relations are classified as semantic relations of core ontology.

³ These tolerance and abbreviation-related terms come from our practice observation.

Semantic relations are classified by two phases: first, we classify BT and NT/RT into the semantic categories; then, we classify the BT-NT/RT relation to certain semantic relations of which domain contains the category of NT and of which range contains the category of BT.

As mentioned in Section 2, the semantic relations with domain and range constraints can be considered relation patterns. With the BT and NT/RT already classified as described in Section 3, the given BT-NT/RT relations can be classified to semantic relations by adopting relation patterns. For example, to the given BT-NT/RT relation of *btnt*(bubble chambers, particle track visualisation), the NT “bubble chambers” can be classified under the category *Equipment*, and the BT “particle track visualization” can be classified under the category *Processing*. Considering that the specific semantic relation of *btnt*(*Equipment*, *Processing*) can only be “*equipmentFor*” in the relation patterns (Figure 4), the given relation is classified as *equipmentFor*(bubble chambers, particle track visualization).

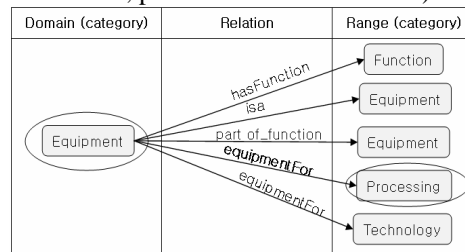


Fig. 4. Semantic relations used as patterns

4.3 Supervised Approach for Semantic Relation Classification

With the training data accumulated, an MEM-based supervised approach is introduced for supervised classification. Each relation triple is treated as one event (which means one example in training data). Only the lexical information of the BT and NT/RT are adopted in the feature extraction, because it is difficult to find usages for each event that contain both BT and NT/RT in the context. The input for the classifier is a feature vector consisting of the following items as basic features:

- The headwords of both BT and NT/RT.
- If event satisfies the rule of the same headword: if yes, the value is 1; else 0.
- If event satisfies the rule of transitivity: if yes, the value is 1; else 0.
- If event satisfies the rule of tolerance: if yes, the value is 1; else 0.
- If event satisfies the rule of abbreviation: if yes, the value is 1; else 0.

Two extra features are adopted for comparison:

- Category feature: categories of the headwords of both BT and NT/RT (use output of Section 3).
- *Is-a* feature: if it is an *is-a* relation the value is 1; else 0 (use output of Section 4.1).

A kNN-based approach is adopted to extract the most similar events from all to build a training model for each target event. Cosine similarity is used for selecting *k* events similar to the target one.

5 Experiments

Coverage and accuracy are used for the evaluation of concept classification. Coverage is adopted to evaluate how many terms are classified into the semantic categories of core ontology. Accuracy is used to evaluate how many terms are classified into the correct categories they should belong to. The evaluation is performed by human experts. One hundred and eighty headwords with top frequencies in a domain dictionary are used as test data, with the assumption that the category of a term is the same as its headword. Coverage of 78% with about 81% of accuracy were achieved in this experiment.

Relation classification uses the automatic concept classification result without verification by human experts.

5.1 Evaluation on *Is-a* Relations

Based on the rules proposed in Section 4.1, a classification system of *is-a* relations is implemented. The BT-NT/RT relations of the Inspec thesaurus are used as test data. The system identifies either *is-a* or *not is-a* relations. Let R1 be relations classified automatically, and R2 be relations decided by human experts. We have accuracy:

$$Accu = \frac{|R1 \cap R2|}{|R2|} \quad (4)$$

As a baseline, classifying all of the BT-NT/RT relations of Inspec to *is-a* relations by default was taken, and came to 79.59% in our test set. The accuracy of the system tested with 4,999 BT-NT/RT relations was 85.83%, obviously higher than the baseline.

Furthermore, evaluation on the results from experts was performed. Relation classification can be confusing even to the experts, so it is possible that differing opinions on a relation exist among experts. This motivated us to evaluate the consistency of the classification results by experts. Let R3 and R4 be classified relations by different experts. The consistency evaluated by Equation (5) was 86.44% in the test with 2,994 relations. This test result shows that the accuracy of the automatic classification system is compatible to the consistency among experts.

$$Cons = \frac{|R3 \cap R4|}{(|R3| + |R4|) / 2} = \frac{|R3 \cap R4|}{|R3|} \quad (5)$$

5.3 Evaluation on Unsupervised Semantic Relation Classification

The similarity-based system described in Section 4.2 was implemented. Among 12,821 BT-NT/RT relations from the Inspec Thesaurus, we derived 3,307 *not is-a* relations after the classification performed in Section 5.2. These *not is-a* relations were classified as 185 non-taxonomic semantic relations automatically. The coverage evaluated with Equation (6) came to 31.09%, and about 90% accuracy was recorded with Equation (4).

$$Coverage = \frac{|\text{Identified NotISA relations}|}{|\text{NotISA relations}|} \quad (6)$$

5.3 Evaluation on Supervised Semantic Relation Classification

An MEM toolkit is employed for model training in MEM-based classification experiments [4]. As the test set, 10% of BT-NT/RT relations were randomly selected from the Inspect Thesaurus, and the remaining 90% of the relations, which had been classified by unsupervised systems and verified by human experts, were used as training data. From Table 2, we can see that the category information of BT and NT/RT is not helpful to the relation classification, while the kNN-based approach remarkably improves accuracy by filtering noise from training data (the lowest row).

Table 2. Experiment results on supervised semantic relations classification.

Approach	Features	Accuracy
MEM	Basic features	59.61%
MEM	Basic features + category features	58.86%
MEM	Basic features + Is-a features	62.46%
MEM	Basic features + category features + Is-a features	61.71%
MEM+kNN	Basic features + Is-a features	66.12%

The supervised approach can overcome the coverage problem of pattern-based approach. However, the accuracy is relatively low. In the experiment shown in the second row of Table 2 (basic features + category features), the accuracy of classification on *is-a* relations was 89.58%, and the accuracy of the other non-taxonomic relations was 24.19%. Obviously, the low accuracy of non-taxonomic relation classification cause the low accuracy of relation classification.

We performed consistency test on human experts, to evaluate the agreement on the non-taxonomic relation classification among human experts. With a randomly selected test set which contains 90 non-taxonomic relations, the consistency among human experts was 15.87%, while the accuracy of the automatic system on the same data set was 14.44%, which was only slightly lower than the consistency of human experts. As the results shown, even human experts felt difficult to classify the BT-NT/RT relations as more than hundreds of the semantic relations. It imposes that too many types of relations can cause low efficiency in relation classification or relation annotation.

6 Related Work

Using an existing knowledge base such as a thesaurus to build ontology has gathered attention over the past few years. Existing thesaurus or thesaurus-like knowledge bases normally provide generally used terms in certain domains as their vocabulary, and present relations between BT and NT/RT with a hierarchy tree structure. Some of these knowledge bases include constraint information that can be used for prediction or reasoning [5]. But they often provide only part of the information required for ontology building.

Some researches focus on converting an existing knowledge base to ontological expression without creating ontology [6]. The original thesaurus formats are

converted into ontological expressions like RDF or OWL. Case studies on the format of individual thesauruses are required, then a pattern or rule-based conversion can be performed. Other researches address the problem of transforming source knowledge into ontological knowledge by extracting useful information from existing knowledge bases [5, 7-8, 10]. Constraints of ontology are derived from the source code of original logic programs, and then transformed to ontological knowledge.

There are also some researches that remodel the thesaurus to ontology with relation level enrichment [1, 8-9]. Some of them [11, 7] extend the thesaurus by inserting case relations and semantic relations into the taxonomic hierarchy of the thesaurus, where case relations are from existing machine translation systems and dictionaries, and the semantic relations acquired from correlation information extracted from corpus. In other researches [1, 9], BT-NT/RT relations are classified as more specific semantic relations of ontology by rules or patterns defined by human developers.

From the point of the relation classification task to solve, what we proposed in this paper is similar to these researches [9, 1]. In our task, however, the IT domain and the Inspec Thesaurus cover a rather broader domain, and so there is difficulty to defining patterns manually. Especially with more than hundreds of the relation types in a broad domain, pattern-based approach shows limitation in the coverage. To solve the problem, a supervised approach is adopted in our task, and the performance was quite promising. Another difference of our task is, the concept classification is performed in this paper to convert the BT-NT/RT hierarchy in domain thesaurus to the taxonomic hierarchy for ontology, but not just allow the terms which do not have taxonomic hypernyms in the domain thesaurus be top nodes in the ontology.

7 Conclusion

This paper presented a method of concept and relation classification to automatically enrich core ontology, with the domain thesaurus Inspec. A frequency and similarity-based approach is proposed for concept classification, to classify the terms under the BT-NT/RT relations to the semantic categories of domain taxonomic hierarchy, thus a BT-NT/RT hierarchy of the thesaurus is converted to a taxonomic hierarchy for ontology. An unsupervised approach is proposed for relation classification without training data, which includes a rule-based approach for *is-a* relation classification, and a pattern-based approach for non-taxonomic relation classification. With the increase in training data, a supervised approach is then presented for relation classification.

Extensive experiments were performed with our proposed approaches. The results show that our approaches are very promising to enrich core ontology automatically with a domain thesaurus. The classification accuracy of *is-a* relations was comparable to the consistency among human experts, and the accuracy of identifying non-taxonomic semantic relations from *not is-a* relations was also high enough for practical use, although coverage was relatively low compared to the accuracy. The coverage problem was overcome with the supervised approach in this paper.

However, the accuracy of the classification on non-taxonomic relations are still very low. It indicates that using only lexical information of the terms is not enough for relation classification, especially when we have more than hundreds of semantic

relations as target categories. How to exploit context information for relation classification is one of our future works.

Another important work which is ongoing at present is, a taxonomic hierarchy among semantic relations should be built for the core ontology, especially when there are many types of relations. With the relation hierarchy, we can reduce the number of target semantic relations, thus the accuracy of relation classification can be improved.

References

1. Dagobert Soergel, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer and Stephen Katz. Reengineering Thesauri for New Applications: the AGROVOC Example. *Journal of Digital Information*, 4(4), March 2004.
2. Key-Sun Choi, Hee-Sook Bae, Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy , In Proceedings of the Global WordNet Conference, pp. 320~325, 2004.1, Brno, Czech.
3. Inspec v2.0 Getting Started Guide. http://scientific.thomson.com/media/scpdf/inspec_gettingstarted_en.pdf
4. Le Zhang. 2004. Maximum Entropy Toolkit for Python and C++. Available from <http://homepages.inf.ed.ac.uk/s0450736/software/maxent/manual.pdf>
5. D. Sleeman, S. Potter, D. Robertson, and M. Schorlemmer. Ontology Extraction for Distributed Environments. In Proceedings of Workshop on Knowledge Transformations for the Semantic Web (affiliated to ECAI-02), July 2002
6. Mark van Assem, Véronique Malaisé, Alistair Miles, and Guus Schreiber: A Method to Convert Thesauri to SKOS. In Proceedings in the 3rd European Semantic Web Conference, June 2006, pp. 95-109
7. Harith Alani, Position paper: Ontology Construction from Online Ontologies. In Proceedings of the 5th International Semantic Web Conference, November 2006
8. Golbeck, Jennifer, Gilberto Fragoso, Frank Hartel, Jim Hendler, Jim Oberthaler, Bijan Parsia "The National Cancer Institute's Thesaurus and Ontology," *Journal of Web Semantics*, 1(1), December 2003.
9. Asanee Kawtrakul, Aurawan Imsombut, Aree Thunkijjanukit, Dagobert Soergel, Anita Liang, Margherita Sini, Gudrun Johannsen, and Johannes Keizer, Automatic Term Relationship Cleaning and Refinement for AGROVOC, Workshop on The Sixth Agricultural Ontology Service, July 25-28, 2005. Vila Real, Portugal.
10. Wielinga, B., Schreiber, G., Wielemaker, J., & Sandberg, J.A.C. From thesaurus to ontology. In Proceedings of International Conference on Knowledge Capture, Victoria, Canada, October 2001
11. Sin-Jae Kang and Jong-Hyeok Lee, Semi-Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries, and Large Corpora, In Proceedings of ACL 2001 Workshop on Human Language Technology and Knowledge Management, Toulouse, France, July 6-7, 2001