Measuring the Specificity of Terms for Automatic Hierarchy Construction

Pum-Mo Ryu[1] and **Key-Sun Choi**[2]

Abstract. This paper introduces new specificity measuring methods of terms using compositional and compositional information. Specificity of a term is the quantity of domain specific information contained in the term. Specific terms have larger quantity of domain information than general terms. Specificity is an important necessary condition for building hierarchical relations among terms. If X_1 is a descendant of X_2, then the specificity of X_1 is greater than that of X_2. As domain specific terms are commonly compounds of the generic level term and some modifiers, compositional information is important to represent the meaning of terms. Contextual information is also used to mitigate the shortcomings of compositional information. Because information theory constitutes a well known formalism for describing information, we adopt the mechanism to measure the information quantity of terms. As the proposed methods do not use domain specific information, they can be applied to other domains without extra processes. Experiments showed very promising results with a precision of 82.0% when applied to terms in the MeSH thesaurus.

1 INTRODUCTION

As current society develops rapidly, new special domains are emerging, and the properties of existing domains are changing continuously. Currently, most domain knowledge is managed manually by domain experts. Manual management is somewhat problematic in terms of coping with the aforementioned rapid changes. As such, automatic domain knowledge management has been focused upon as a new research area. As terms are linguistic realizations of domain specific concepts, term management is a core part of domain knowledge management [1]. Constructing hierarchical relations among terms is one of the major tasks in term management and is used in various applications, including information retrieval, document classification, information extraction, and knowledge representation.

Specificity of a term is the quantity of domain specific information contained in the term. Some terms have a relatively large quantity of domain information, and others have a relatively small quantity of domain information. Specific terms cover narrow range in conceptual level and tend to locate at deep level in term hierarchy. Because the specificity of descendent terms is higher than that of the ancestor terms, specificity is a kind of necessary condition for hierarchical structure; *i.e.* if X_1 is an ancestor of X_2, then the specificity of X_1 is less than the specificity of X_2. However, this is not a satisfactory condition for hierarchical structure. When the specificity of X_1 is less than the specificity of X_2, X_1 is not always an ancestor of X_2. Because it is very difficult to determine satisfactory conditions for term hierarchy, several necessary conditions such as specificity of terms and similarity among terms are used instead. Many works on automatic construction of term hierarchy were based on similarity measures. To date, well formalized specificity measuring methods have not been proposed, and specificity of terms is not widely used in this research area.

When domain specific concepts are represented as terms, the terms are classified into two categories based on the composition of part words. In the first category, new terms are created by adding modifiers to existing terms. For example "*insulin-dependent diabetes mellitus*" was created by adding the modifier "*insulin-dependent*" to its hypernym "*diabetes mellitus*", as in Table 1. In English, the specific level terms are commonly compounds of the generic level term and some modifier [2]. In this case, compositional information is important to represent their information. In the second category, new terms are created independently of existing terms. For example, although "*wolfram syndrome*" is semantically related to its ancestor terms, it shares no common words with its ancestor terms in Table 1. In this case, contextual information is used to discriminate the features of the terms. Contextual information is also used to address additional semantic components that cannot be predicted from the parts of terms. Therefore, our specificity measuring methods are based on both compositional and contextual information of terms. In this paper, we propose new specificity measuring methods based on information theory. By the theory, specificity is quantified to a positive real number as given in equation (1.1).

$$Spec(X) \in R^{+} \tag{1.1}$$

where X is a term, and *Spec(X)* is the specificity of X.

Table 1. Subtree of the MESH[3] tree. Numbers represent hierarchical structure of terms

Node Number	Terms
C18.452.297	diabetes mellitus
C18.452.297.267	insulin-dependent diabetes mellitus
C18.452.297.267.960	wolfram syndrome

1,2 KAIST/KORTERM/BOLA, Daejeon, Korea
email: {pmryu,kschoi}@world.kaist.ac.kr

3 MeSH is available at http://www.nlm.nih.gov/mesh. MeSH 2003 was used in this research.

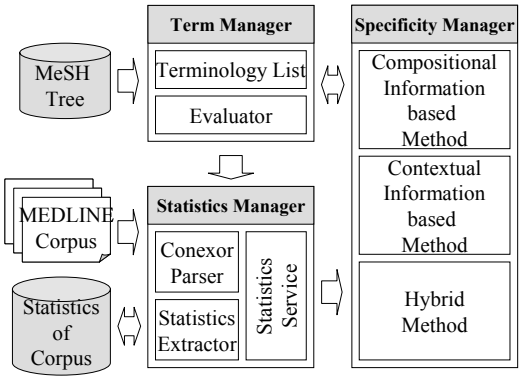

Figure 1. System overview

The system is composed of three managers as Figure 1, Term manager, Statistics manager, and Specificity manager. The roles of these managers are as follows:

- Term Manager manages a term list, which is a target of specificity calculation. This manager also evaluates the specificity values calculated by various methods.
- Statistics Manager extracts and stores various statistics from the corpus, and provides the statistics to the specificity manager.
- Specificity Manager provides various specificity calculation functions described in this paper.

The remainder of this paper is organized as follows. Related works are introduced in Section 2, characteristics of compositional and contextual information are described in Section 3, the specificity measuring methods based on information theory are introduced in Section 4, and experiments and evaluation on the methods are discussed in Section 5. Finally, conclusions are drawn in Section 6.

2 RELATED WORKS

In this section, we describe previous works on automatic construction of hyponymy relations. We also describe a previous work on specificity measuring methods for nouns.

There are two main approaches to extract hyponymy relations from a corpus; the first class of approaches is based on lexico-syntactic patterns, and the second on the contextual similarity of words.

Hearst [3] and Caraballo [4] extracted hyponymy relations using lexico-syntactic patterns. This method can be applied to all domains without additional information, and can extract relatively accurate relations. However, this method extracts hyponymy relations that are explicitly connected by patterns.

Pereira [5] clustered nouns based on distributions of predicates that have the nouns as direct objects. Deterministic annealing is used to make a hierarchical structure of nouns in a top-down manner. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical structure. Grefenstette [6] developed the SEXTANT system to make clusters of semantically related words. He constructed contextual information of words using extracted patterns of adjective-noun, subject-predicate, and object-predicate. The contextual information was compared using a weighted Jaccard similarity measure to determine similar words. Sanderson [7] used the concept of subsumption between contexts to create a hierarchical structure of terms. In this work, if two terms, X and Y, satisfy equation (2.1) then X is a hypernym of Y.

$$p(X \mid Y) = 1,\ p(Y \mid X) < 1 \tag{2.1}$$

where $P(X|Y)$ is the probability of Y occurring when X appears in the corpus. The methods described in this section have been mainly applied to general level words, and contextual information was used to represent the meaning of the words. However, domain specific terms have sufficient information in themselves. Therefore, compositional information of terms is also important to make hyponymy relations among terms.

Caraballo [8] calculated the specificity of nouns using modifier distributions. The basic assumption of this research is that general nouns are frequently modified by modifiers, and specific nouns are rarely modified by modifiers. They calculated entropy of the rightmost prenominal modifier for a noun, and compared the entropy values to determine specificity. As general nouns have a complex modifier distribution, the entropies for these nouns are high. For terms, as sufficient modifiers are not extracted from the corpus, this method has a limitation of data sparseness.

3 CHARACTERISTICS OF COMPOSITIONAL AND CONTEXTUAL INFORMATION

In this section, we describe compositional information and contextual information which is used to measure the specificity of terms.

Domain specific concepts have their own feature sets. More specific concepts are created by adding other features to the feature set of existing concepts. Let's consider two concepts X and Y. X is an existing concept and Y is a newly created concept by adding new features to X. In this case, X is a hyper concept of Y, and the feature set of X is a subset of that of Y [9]. When Y is mapped to a term, the term may or may not share common words with the term for X. If a new term shares many common words with its ancestors, compositionality is important to measure the relative specificity of terms. Otherwise, contextual information is important.

3.1 Compositional information

This section describes compositionality of terms. By compositionality, the meaning of the whole term can be strictly predicted from the meaning of the individual words. Many terms are created by appending modifiers to existing terms. In this mechanism, features of modifiers are added to the feature set of existing terms to make new concepts. All unit words in a term have their own features, and their features are summed up to make a feature set of the original term. On this assumption, we use word frequency, tf.idf, bigram to quantify features of unit words.

Word frequency has been major criteria in automatic term recognition (ATR) [10,11]. In ATR works, high frequency candidates are highly probable to be domain terms. Contrarily, we assume that terms having low frequency words obtain a high specificity value. Because low frequency words appear only in a small number of terms, the words can clearly discriminate the

terms from other terms. The difference between ATR works and our work is that the former use frequency of terms whereas our research uses the frequency of unit words composing terms.

tf.idf, a multiplied value of term frequency (tf) and inverse document frequency (idf), is a widely used term weighting scheme in information retrieval [12], as given by equation (3.1).

$$tf \cdot idf(w) = \begin{cases} (1+\log tf(w))\log(\frac{N}{df(w)}) & if \quad tf(w) \geq 1 \\ 0 & if \quad tf(w) = 0 \end{cases} \quad (3.1)$$

where *tf(w)*, *df(w)*, *tf.idf(w)* are term frequency, document frequency, and tf.idf of word *w*, respectively, and *N* is the number of documents. Words with high term frequency and low document frequency receive a high tf.idf value. Because a document usually discusses one topic, and words with high tf.idf values are good index terms for the document, the words are considered to have topic specific information. Therefore, if a term has words with high tf.idf values, the term is assumed to have topic specific information.

Bigram is also useful information to quantify meaning of unit words. In a bigram based method, unit words in a term are only dependent on previously adjacent words. If a term has bigrams of low frequency, then the term receives high specificity values. As bigrams of low frequency appear in a small number of terms, they are highly capable of distinguishing the terms from other terms. If the bigram based method shows better results than the word or tf.idf based method, chunks of words can be considered to be more informative than independent words.

The internal structure of terms is also important information for term specificity. We assume that many terms are compound nouns made of one or more modifiers and one head noun. If the modifier-head structure of a term is known, the specificity of the term is calculated incrementally starting from the head noun. In this manner, the specificity value of a term is always larger than that of the head term. This result is consistent with the assumption that specific terms have larger specificity values. However, it is very difficult to analyze the modifier-head structure of a compound noun. We use simple nesting relations between terms to analyze the structure of terms. A term *X* is nested in term *Y*, when *X* is a substring of *Y* [10]. A more detailed definition is as follows:

Definition 1 If two terms *X* and *Y* are terms in the same category and *X* is nested in *Y* as W_1XW_2, then *X* is a head term, and W_1 and W_2 are modifiers of *X*.

For example two terms, "*diabetes mellitus*" and "*insulin dependent diabetes mellitus*", are both disease names, and the former is nested in the latter. In this case, "*diabetes mellitus*" is the head term and "*insulin dependent*" is a modifier of "*insulin dependent diabetes mellitus*" by definition 1. If multiple terms are nested in a term, the longest term is selected as the head term. The specificity of *Y* is measured as equation (3.2).

$$Spec(Y) = Spec(X) + \alpha \cdot Spec(W_1) + \beta \cdot Spec(W_2) \quad (3.2)$$

where *Spec(X)*, $Spec(W_1)$, and $Spec(W_2)$ are specificity values of *X*, W_1, and W_2, respectively. α and β, real numbers between 0 and 1, are weighting schemes for the specificity of modifiers, and are obtained experimentally.

3.2 Contextual information

In this section, contextual information is introduced to overcome problems of compositional information.

There are problems that cannot be addressed by the compositionality of terms. First, let's consider again the three disease names in Table 1. Although the feature set of "*wolfram syndrome*" shares many common features with the feature set of "*insulin-dependent diabetes mellitus*" at a semantic level, they do not share common words at a lexical level. In this case, it is undesirable to compare two specificity values measured using compositional information alone. Second, when several words are combined in a term, there are additional semantic components that are not predicted by unit words. For example, "*wolfram syndrome*" is a kind of "*diabetes mellitus*" The meaning "*diabetes mellitus*" is not predicted by two separate words "*wolfram*" and "*syndrome*". Finally, in some instances the modifier-head structure of a term is not determined by the compositional information. For instance, "*vampire slayer*" might be a slayer who is vampire or a slayer of vampires. Contextual information can complement the compositional information and thereby mitigate this problem.

Contextual information is the distribution of surrounding words of target terms. For example, the distribution of co-occurrence words of the terms, the distribution of predicates that have the terms as arguments, and the distribution of modifiers of the terms comprise contextual information.

General terms tend to be modified by other words. Conversely, domain specific terms do not tend to be modified by other words, because they have sufficient information in themselves [8]. Under this assumption, we use the probabilistic distribution of modifiers as contextual information. Because domain specific terms, unlike general words, are rarely modified in a corpus, it is important to collect sufficient modifiers from a given corpus. Therefore accurate text processing such as syntactic parsing is needed to extract modifiers. For general words Caraballo [8] extracted only rightmost prenominals as context information. We use the Conexor functional dependency parser for English [13] to analyze the structure of sentences. Among many dependency functions defined in the Conexor parser, *attr* and *mod* functions are used to extract modifiers. In this manner, more plentiful modifiers are extracted than in previous research. If a term or modifiers of the term do not occur in the corpus, the specificity of the term cannot be measured using contextual information.

4 SPECIFICITY MEASURING METHODS

In this section, we describe information theory *like* specificity measuring methods using compositional and contextual information. Here, we call information theory like methods, because some probability values used in these methods are not real probabilities; rather they are the relative weights of terms of words. Because information theory constitutes a well known formalism for describing information, we adopt the mechanism to measure the information quantity of terms.

In information theory, when a low probability message occurs on a channel output, the amount of *surprise* is large, and the length of bits to represent this message becomes long. Therefore, a large quantity of information is gained by this message [14]. If we consider the terms in a corpus as messages of a channel output, the information quantity of the terms can be measured

using various statistics acquired from the corpus. A set of terms is defined as equation (4.1) for further explanation.

$$T = \{t_k \mid 1 \le k \le n\} \tag{4.1}$$

where t_k is a term and n is the total number of terms. In the next step, a discrete random variable X is defined as equation (4.2).

$$\begin{aligned} X &= \{x_k \mid 1 \le k \le n\} \\ p(x_k) &= \text{Prob}(X = x_k) \end{aligned} \tag{4.2}$$

where x_k is an event of term t_k occurring in the corpus, $p(x_k)$ is the probability of event x_k. The information quantity, $I(x_k)$, gained after observing event x_k, is defined by a logarithmic function. Finally, $I(x_k)$ is used as specificity value of t_k as equation (4.3).

$$Spec(t_k) \approx I(x_k) = -\log p(x_k) \tag{4.3}$$

In equation (4.3), if we can estimate the probability of event x_k, $p(x_k)$, then we can measure the specificity value of t_k. We describe some estimating methods of $p(x_k)$ in the following sections.

4.1 Compositional information based method (method 1)

In this section, we describe a specificity measuring method using compositional information, introduced in section 3.1. This method is divided into two steps: In the first step, specificity values of all words are measured independently. In the second step, the specificity values of words are summed up. For a detailed description, we assume that a term t_k consists of one or more words as given by equation (4.4).

$$t_k = w_1 w_2 ... w_m \tag{4.4}$$

where w_i is a word in t_k. In the next step, a discrete random variable Y is defined as equation (4.5).

$$\begin{aligned} Y &= \{y_i \mid 1 \le i \le m\} \\ p(y_i) &= \text{Prob}(Y = y_i) \end{aligned} \tag{4.5}$$

where y_i is an event of a word w_i occurring in the corpus, and $p(y_i)$ is the probability of event y_i. Information quantity $I(x_k)$ in equation (4.3) is redefined as equation (4.6) based on the previous assumption.

$$I(x_k) = -\sum_{i=1}^{m} p(y_i) \log p(y_i) \tag{4.6}$$

where $p(y_i)$ is the probability of the event that word w_i occurs in the corpus, and $I(x_k)$ is the average information quantity of all words in t_k. Three types of information, word frequency, tf.idf, and bigram, are used to estimate $p(y_i)$. In this mechanism, the probability for high informative words should be low.

When word frequency is used to quantify features of words, $p(y_i)$ in equation (4.6) is estimated as equation (4.7).

$$p(y_i) \approx p_{MLE}(w_i) = \frac{freq(w_i)}{\sum_j freq(w_j)} \tag{4.7}$$

where $freq(w)$ is the frequency of word w in the corpus, $P_{MLE}(w_i)$ is the maximum likelihood estimation of $P(w_i)$, and j is an index of all words in the corpus. In this equation, $P(w_i)$ for low frequency words is high.

When tf.idf is used to quantify features of words, $p(y_i)$ in equation (4.6) is estimated as equation (4.8).

$$p(y_i) \approx p_{MLE}(w_i) = 1 - \frac{tf \cdot idf(w_i)}{\sum_j tf \cdot idf(w_j)} \tag{4.8}$$

where $tf{\cdot}idf(w)$ is the tf.idf value of word w. In this equation, as large tf.idf words are informative, $p(y_i)$ of the words becomes low.

When a bigram is used to quantify features of words, $p(y_i)$ in equation (4.6) is estimated as equation (4.9).

$$p(y_i) \approx \begin{cases} p_{MLE}(w_i) = \dfrac{freq(w_i)}{\sum_j freq(w_j)} & \text{if } i = 1 \\ p_{MLE}(w_i \mid w_{i-1}) = \dfrac{freq(w_{i-1} w_i)}{\sum_j freq(w_{i-1} w_j)} & \text{if } i > 1 \end{cases} \tag{4.9}$$

where $P_{MLE}(w_1)$ is the probability of the first word w_1, and $P_{MLE}(w_i|w_{i-1})$ is the probability of word w_i being located at position i, when word w_{i-1} is located at position $i-1$ in t_k. $freq(w_1 w_2)$ is the frequency of w_1 and w_2 occurring adjacently in the corpus.

4.2 Contextual information based method (method 2)

In this section, we describe a method using contextual information, introduced in section 3.2.

Entropy of probabilistic distribution of modifiers for a term is defined as equation (4.10).

$$H_{mod}(t_k) = -\sum_i p(mod_i, t_k) \log p(mod_i, t_k) \tag{4.10}$$

where $p(mod_i, t_k)$ is the probability that mod_i modifies t_k and is estimated as equation (4.11).

$$p_{MLE}(mod_i, t_k) = \frac{freq(mod_i, t_k)}{\sum_j freq(mod_j, t_k)} \tag{4.11}$$

where $freq(mod_i, t_k)$ is the frequencies with which mod_i modifies t_k in corpus, j is index of all modifiers of t_k in corpus. The entropy calculated by equation (4.10) is the average information quantity of all (mod_i, t_k) pairs. Specific terms have low entropy, as their modifier distributions are simple. Therefore inversed entropy is assigned to $I(x_k)$ in equation (4.3) to allow specific terms to obtain large information quantity, as delineated in equation (4.12).

$$I(x_k) \approx \max_{1 \le i \le n} (H_{mod}(t_i)) - H_{mod}(t_k) \tag{4.12}$$

where the first term of approximation is the maximum modifier entropy among the entropies of all terms.

4.3 Hybrid method (method 3)

There are some shortcomings in the previous two methods. Comparing specificity values measured by compositional

Table 2. The specificity values of terms were measured with method 1, method 2, and method 3. In method 1, the word frequency based method, the word tf.idf based method, the word bigram based method, and structure information added methods were separately experimented. Two additional methods, based on term frequency and term tf.idf, were experimented. The two methods displaying the best performance in method 1 and method 2 were combined into method 3.

Methods		Precision (%)			Coverage (%)
		Type I	Type II	Total	
Human subjects(Average)		96.6	86.4	87.4	
Term frequency		100.0	53.5	60.6	89.5
Term tf·idf		52.6	59.2	58.2	89.5
Compositional Information Method (Method 1)	Word frequency	37.2	72.5	69.0	100.0
	Word frequency + Structure (α=β=0.2)	100.0	72.8	75.5	100.0
	Word tf·idf	44.2	75.3	72.2	100.0
	Word tf·idf +Structure (α=β=0.2)	100.0	76.6	78.9	100.0
	Word Bigram	37.2	59.5	57.3	100.0
	Word Bigram + Structure (α=β=0.3)	100.0	60.6	64.4	100.0
Contextual Information Method (mod cnt>1) (Method 2)		90.0	66.4	70.0	70.2
Hybrid Method (tf·idf + Structure, γ=0.8) (Method 3)		95.0	79.6	82.0	70.2

information methods is inappropriate if two terms do not share common words. Furthermore, specificity values of terms cannot be measured by a contextual information based method if the terms or modifiers of the terms do not occur in the corpus. To overcome these shortcomings, a hybrid method is introduced, as delineated by equation (4.13).

$$I(x_k) \approx \frac{1}{\gamma(\frac{1}{I_{Cmp}(x_k)}) + (1-\gamma)(\frac{1}{I_{Ctx}(x_k)})} \tag{4.13}$$

where $I_{Cmp}(x_k)$ and $I_{Ctx}(x_k)$ are normalized $I(x_k)$ values between 0 and 1, which are calculated by the compositional and contextual information based methods, respectively. $\gamma(0 \le \gamma \le 1)$ is the weight of two values. If $\gamma = 0.5$, the equation is the harmonic mean of the two values. Therefore, $I(x_k)$ becomes large when the two values are equally large.

5 EXPERIMENTS AND EVALUATION

In this section, we describe experiments and evaluate the proposed methods. For convenience, we simply refer to the compositional information based method, contextual information based method, and hybrid method as method 1, method 2, and method 3, respectively.

5.1 Experiment

A sub-tree of the MeSH thesaurus is selected for the experiment. "*metabolic diseases(C18.452)*" node is the root of the subtree, and the subtree consists of 436 disease names, which are target terms of specificity measuring. A set of journal abstracts was extracted from the MEDLINE[4] database using the disease names as search queries. Therefore, all the abstracts are related to some of the disease names in the subtree. The set consists of approximately 170,000 abstracts (20,000,000 words). The abstracts are analyzed by the Conexor parser, and various statistics are extracted: 1) frequency, tf.idf of the disease names; 2) distribution of modifiers of the disease names; 3) frequency, tf.idf of unit words of the disease names; and 4) unit word bigrams of the disease names.

We divided parent-child relations into two types. Relations in which parent terms are nested in child terms are categorized as type I. Other relations are categorized as type II. There are 43 type I relations in and 393 type II relations. The type I relations always have correct specificity values if the structural information method described in section 3.1 is applied.

We tested 10 human subjects to find the upper bound of precision. The subjects were all medical doctors of internal medicine, a tightly related division to "*metabolic diseases*". They were asked to identify the parent-child relation of given two terms. The average precisions of type I and type II were 96.6% and 86.4%, respectively. We set these values as the upper bound of precision. The precision of type I relations was less than 100%, although the relations are easily identified using simple rules. It is thus assumed that this result reflects mistakes made by the subjects.

The specificity values of terms were measured with method 1, method 2, and method 3, as presented in Table 2. In method 1, the word frequency based method, the word tf.idf based method, the word bigram based method, and structure information added methods were separately experimented. Two additional methods, based on term frequency and term tf.idf, were experimented to compare the contribution of the terms themselves and that of the unit words composing the terms. The two methods displaying the best performance in method 1 and method 2 were combined into method 3.

5.2 Evaluation

The system was evaluated by two criteria, coverage and precision. Coverage is the fraction of the terms that have specificity values by the given measuring method, as given by equation (5.1).

4 MEDLINE is a database of biomedical articles serviced by National Library of Medicine, USA. (http://www.nlm.nih.gov)

$$coverage = \frac{\#\ of\ terms\ with\ specificity}{\#\ of\ all\ terms} \quad (5.1)$$

Method 2 obtains relatively lower coverage than method 1, because method 2 can measure specificity only when the terms and their modifiers appear in the corpus. On the contrary, method 1 can measure the specificity of the terms when parts of unit words appear in the corpus. Precision is the fraction of relations with correct specificity values, as given by equation (5.2).

$$precision = \frac{\#\ of\ R(p,c)\ with\ correct\ specificity}{\#\ of\ all\ R(p,c)} \quad (5.2)$$

where *R(p,c)* is a parent-child relation in the MeSH thesaurus, and the relation is valid when specificity of two terms are measured by the given method. If the specificity value of child term *c* is larger than that of parent term *p*, then the relation has correct specificity values.

The word frequency and tf.idf based methods showed better performance than the word bigram based method and term based methods. This result indicates that the information of terms tends to be divided into unit words independently rather than into whole terms. This result also illustrates the basic assumption of this paper that specific concepts are created by adding information to existing concepts, and new concepts are expressed as new terms by adding modifiers to existing terms. The word tf.idf based method showed better performance than the word frequency based method in method 1. This result indicates that word tf.idf is more informative than word frequency.

Method 2 showed the best performance, a precision of 70.0% and coverage of 70.2%, when we selected modifiers that modify target terms two or more times. However, method 2 showed worse performance than the word tf.idf and structure based methods. It is assumed that because domain specific terms are rarely modified by other words, sufficient contextual information was not extracted from the corpus.

Method 3, a hybrid method of method 1 (tf.idf of words, structure information) and method 2, showed the best precision, i.e., 82.0%, because the two methods interacted in a complementary manner in the process. The coverage of this method was 70.2%, which equals that of method 2, because the hybrid specificity value is calculated only when the specificity of method 2 is valid. In the hybrid method, the weight value $\gamma = 0.8$ indicates that compositional information is more important than contextual information when measuring the specificity of domain-specific terms. The precision of 82.0% reflects good performance relative to the upper bound of 87.4%.

One reason of the errors is that the names of some internal nodes in the MeSH thesaurus are category names rather than disease names. For example, as "*acid-base imbalance (C18.452.076)*" is the name of a disease category, it does not occur as frequently as other real disease names. Other cause of the error is that we did not consider various surface forms of the same term. For example, although "*NIDDM*" is the acronym of "*non insulin dependent diabetes mellitus*", the system counted two terms independently. Therefore the extracted statistics do not properly reflect the semantic level information.

If we analyze the morphological structure of terms, some problems can be solved. For example, "*nephrocalcinosis*" has a modifier-head structure at the morpheme level: "*nephro*" is the modifier and "*calcinosis*" is the head. Therefore, if an internal word analysis is possible, the internal structure method of section 3.1 also can be applied at the morpheme level. As word formation rules are heavily dependent on the domain specific morphemes, additional analysis is needed to apply this approach to other domains.

6 CONCLUSIONS

This paper proposed specificity measuring methods for terms based on information theoretic measures using the compositional and contextual information of terms. Specificity of terms is an important necessary condition to make term hierarchy automatically. The methods are experimented on terms from the MeSH thesaurus. The proposed hybrid method showed the best precision of 82.0%, as the two methods mitigated each other's respective shortcomings. As the proposed methods do not use domain dependent information, the methods can easily be adapted to other domains.

In the future, the system will be modified to handle various term formations such as abbreviated forms. Morphological structure analysis of words is also needed to use the morpheme level information.

ACKNOWLEDGEMENTS

This work was supported in part by the Ministry of Science & Technology, the Ministry of Culture & Tourism of the Korean government, and the Korea Science & Engineering Foundation. We also would like to thank the members of the Internal Medicine Department of the Daegu Catholic Univ. Medical Center, Korea.

REFERENCES

[1] J. C. Sager, *Handbook of Term Management: volume 1*, John Benjamins publishing company (1997)

[2] William Croft, *Typology and Universals 2nd edition*, Cambridge Textbooks in Linguistics, Cambridge Univ. Press (2004)

[3] Marti A. Hearst, *Automatic Acquisition of Hyponyms from Large Text Corpora*, In the proceedings of ACL (1992)

[4] S. A. Caraballo, *Automatic construction of a hypernym-labeled noun hierarchy from text Corpora*, In the proceedings of ACL (1999)

[5] Fernando Pereira, Naftali Tishby, and Lillian Lee, *Distributational clustering of English words*, In the proceedings of ACL (1993)

[6] Gregory Grefenstette, *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers (1994)

[7] Mark Sanderson, *Deriving concept hierarchies from text*, In the proceedings of the 22th Annual ACM S1GIR Conference on Research and Development in Information Retrieval (1999)

[8] S. A. Caraballo and E. Charniak, *Determining the Specificity of Nouns from Text*, In the proceedings of the Joint SIGDAT Conference on EMNLP and Very Large Corpora (1999)

[9] ISO 704, *Terminology work-Principles and methods*, ISO 704:2000(E) (2000)

[10] Katerina Frantzi, Sophia Anahiadou and Hideki Mima, *Automatic recognition of multi-word terms: the C-value/NC-value method*, Journal of Digital Libraries, Vol. 3, Num. 2 (2000)

[11] Jong-Hoon Oh, Kyung-Soon Lee and Key-Sun Choi, *Term Recognition Using Technical Dictionary Hierarchy*, In the proceedings of ACL (2000)

[12] Christopher D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press (1999)

[13] Conexor, *Conexor: Functional Dependency Grammar Parser*, http://www.conexor.fi (2004)

[14] Simon Haykin, *Neural Network*, IEEE Press, pp. 444 (1994)