Determining the Specificity of Terms based on Information Theoretic Measures

Pum-Mo Ryu and Key-Sun Choi
Dept. EECS/KORTERM KAIST
373-1 Guseong-dong Yuseong-gu
305-701 Daejeon
Korea
pmryu@world.kaist.ac.kr, kschoi@world.kaist.ac.kr

Abstract
This paper introduces new specificity determining methods for terms based on information theoretic measures. The specificity of terms represents the quantity of domain specific information that is contained in the terms. Compositional and contextual information of terms are used in proposed methods. As the methods don’t rely on domain dependent information, they can be applied to other domains without extra processes. Experiments showed very promising results with the precision 82.0% when applied to the terms in MeSH thesaurus.

1 Introduction
The specificity of terms represents the quantity of domain specific information contained in the terms. If a term has large quantity of domain specific information, the specificity of the term is high. The specificity of a term \( X \) is quantified to positive real number as equation (1).

\[
\text{Spec}(X) \in \mathbb{R}^+
\]  

(1)

The specificity is a kind of necessary condition for term hierarchy, i.e., if \( X_1 \) is one of ancestors of \( X_2 \), then \( \text{Spec}(X_1) \) is less than \( \text{Spec}(X_2) \). Thus this condition can be applied to automatic construction or evaluation of term hierarchy. The specificity also can be applied to automatic term recognition.

Many domain specific terms are multiword terms. When domain specific concepts are represented as multiword terms, the terms are classified into two categories based on composition of unit 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 modifier “insulin-dependent” to its hypernym “diabetes mellitus” as in Table 1. In English, the specific level terms are very commonly compounds of the generic level term and some modifier (Croft, 2004). In this case, compositional information is important to get meaning of the terms. In the second category, new terms are independent of existing terms. For example, “wolfram syndrome” is semantically related to its ancestor terms as in Table 1. But it shares no common words with its ancestor terms. In this case, contextual information is important to get meaning of the terms.

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

Table 1 Subtree of MeSH\(^1\) thesaurus. Node numbers represent hierarchical structure of terms

Contextual information has been mainly used to represent the meaning of terms in previous works. (Grefenstette, 1994) (Pereira, 1993) and (Sanderson, 1999) used contextual information to find hyponymy relation between terms. (Caraballo, 1999) also used contextual information to determine the specificity of nouns. Contrary, compositional information of terms has not been commonly discussed. We propose new specificity measuring methods based on both compositional and contextual information. The methods are formulated as information theory like measures.

This paper consists as follow; new specificity measuring methods are introduced in section 2, and the experiments and evaluation on the methods are discussed in section 3, finally conclusions are drawn in section 4.

2 Specificity Measuring Methods
In this section, we describe information theory like methods to measure the specificity of terms. Here, we call information theory like methods, because some probability values used in these methods are

\(^1\) MeSH is available at http://www.nlm.nih.gov/mesh. MeSH 2003 was used in this research.
not real probability, rather they are relative weight of terms or words.

In information theory, when a low probability message occurs on channel output, the quantity of surprise is large, and the length of bits to represent the message becomes long. Thus the large quantity of information is gained by the message (Haykin, 1994). If we regard the terms in corpus as the messages of channel output, the information quantity of the terms can be measured using information theory. A set of target terms is defined as equation (2) for further explanation.

\[ T = \{ t_k | 1 \leq k \leq n \} \]  
(2)  

where \( t_k \) is a term. In next step, a discrete random variable \( X \) is defined as equation (3).

\[ X = \{ x_k | 1 \leq k \leq n \}, \quad p(x_k) = \text{Prob}(X = x_k) \]  
(3)  

where \( x_k \) is an event of \( t_k \) is observed in corpus, \( p(x_k) \) is the probability of \( x_k \). The information quantity, \( I(x_k) \), gained after observing \( x_k \), is used as the specificity of \( t_k \) as equation (4).

\[ \text{Spec}(t_k) \approx I(x_k) = -\log p(x_k) \]  
(4)  

By equation (4), we can measure the specificity of \( t_k \) by estimating \( p(x_k) \). We describe three estimating methods for \( p(x_k) \) in following sections.

### 2.1 Compositional Information based Method (Method 1)

By compositionality, the meaning of a term can be strictly predicted from the meaning of the individual words (Manning, 1999). This method is divided into two steps: In the first step, the specificity of each word is measured independently. In the second step, the specificity of composite words is summed up. For detail description, we assume that \( t_k \) consists of one or more words as equation (5).

\[ t_k = w_i \cdot w_{i+1} \ldots w_n \]  
(5)  

where \( w_i \) is \( i \)-th word in \( t_k \). In next step, a discrete random variable \( Y \) is defined as equation (6).

\[ Y = \{ y_i | 1 \leq i \leq n \}, \quad p(y_i) = \text{Prob}(Y = y_i) \]  
(6)  

where \( y_i \) is an event of \( w_i \) occurs in term \( t_k \), \( p(y_i) \) is the probability of \( y_i \). Information quantity, \( I(x_k) \), in equation (4) is redefined as equation (7) based on previous assumption.

\[ I(x_k) = -\sum_{j=1}^{n} p(y_j) \log p(y_j) \]  
(7)  

where \( I(x_k) \) is average information quantity of all words in \( t_k \). In this mechanism, \( p(y_j) \) of informative words should be smaller than that of non-informative words. Two information sources, word frequency, \( \text{tf.idf} \) are used to estimate \( p(y_j) \) independently.

We assume that if a term is composed of low frequency words, the term have large quantity of domain information. Because low frequency words appear in limited number of terms, they have high discriminating ability. On this assumption, \( p(y_j) \) in equation (7) is estimated as relative frequency of \( w_i \) in corpus. In this estimation, \( P(y_j) \) for low frequency words becomes small.

\( \text{tf.idf} \) is widely used term weighting scheme in information retrieval (Manning, 1999). We assume that if a term is composed of high \( \text{tf.idf} \) words, the term have domain specific information. On this assumption, \( p(y_j) \) in equation (7) is estimated as equation (8).

\[ p(y_j) \approx \text{pMLE}(w_i) = 1 - \frac{\text{tf.idf}(w_i)}{\sum_j \text{tf.idf}(w_j)} \]  
(8)  

where \( \text{tf.idf}(w) \) is \( \text{tf.idf} \) value of \( w \). In this equation, \( p(y_j) \) of high \( \text{tf.idf} \) words becomes small.

If the modifier-head structure is known, the specificity of the term is calculated incrementally starting from head noun. In this manner, the specificity of the term is always larger than that of the head term. This result answers to the assumption that more specific term has higher specificity. We use simple nesting relations between terms to analyze modifier-head structure as follows (Frantzi, 2000):

**Definition 1** If two terms \( X \) and \( Y \) are terms in same semantic category and \( X \) is nested in \( Y \) as \( W_1 \cdot W_2 \), then \( X \) is head term, and \( W_1 \) and \( W_2 \) are modifiers of \( X \).

For example, because “\textit{diabetes mellitus}” is nested in “\textit{insulin dependent diabetes mellitus}” and two terms are all disease names, “\textit{diabetes mellitus}” is head term and “\textit{insulin dependent}” is modifier. The specificity of \( Y \) is measured as equation (9).

\[ \text{Spec}(Y) = \text{Spec}(X) + \alpha \cdot \text{Spec}(W_1) + \beta \cdot \text{Spec}(W_2) \]  
(9)  

where \( \text{Spec}(X) \), \( \text{Spec}(W_1) \), and \( \text{Spec}(W_2) \) are the specificity of \( X \), \( W_1 \), \( W_2 \) respectively. \( \alpha \) and \( \beta \) are weighting schemes for the specificity of modifiers. They are found by experimentally.

### 2.2 Contextual Information based Method (Method 2)

There are some problems that are hard to address using compositional information alone. Firstly, although two disease names, “\textit{wolfram syndrome}” and “\textit{insulin-dependent diabetes mellitus}”, share
many common features in semantic level, they don’t share any common words in lexical level. In this case, it is unreasonable to compare two specificity values based on compositional information. Secondly, when several words are combined into one term, there are additional semantic components that are not predicted by unit words. For example, “wolfram syndrome” is a kind of “diabetes mellitus”. We can not predict the meaning of “diabetes mellitus” from two separate words “wolfram” and “syndrome”. Thus we use contextual information to address these problems.

General terms are frequently modified by other words in corpus. Because domain specific terms have sufficient information in themselves, they are rarely modified by other words, (Caraballo, 1999). Under this assumption, we use probability distribution of modifiers as contextual information. Collecting sufficient modifiers from given corpus is very important in this method. To this end, we use Conexor functional dependency parser (Conexor, 2004) to analyze the structure of sentences. Among many dependency functions defined in the parser, “attr” and “mod” functions are used to extract modifiers from analyzed structures. This method can be applied the terms that are modified by other words in corpus.

Entrophy of modifiers for a term is defined as equation (10).

\[
H_{mod}(t_k) = -\sum_{i} p(mod_i,t_k) \log p(mod_i,t_k)
\]  

(10)

where \( p(mod_i,t_k) \) is the probability of \( mod_i \) modifies \( t_k \) and it is estimated as relative frequency of \( mod_i \) in all modifiers of \( t_k \). The entropy calculated by equation (10) is the average information quantity of all \( (mod_i,t_k) \) pairs. Because domain specific terms have simple modifier distributions, the entropy of the terms is low. Therefore inversed entropy is assigned to \( I(x_k) \) in equation (4) to make specific terms get large quantity of information.

\[
I(x_k) \approx \max_{i \geq 1}(H_{mod}(t_i))-H_{mod}(t_k)
\]

(11)

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

2.3 Hybrid Method (Method 3)

In this section, we describe hybrid method to overcome shortcomings of previous two methods. In this method the specificity is measured as equation (12).

\[
I(x_k) \approx \frac{1}{\gamma \left( \frac{1}{I_{Cmp}(x_k)} \right) + (1-\gamma)\left( \frac{1}{I_{Ctx}(x_k)} \right)}
\]

(12)

where \( I_{Cmp}(x_k) \) and \( I_{Ctx}(x_k) \) are information quantity measured by method1 and method 2 respectively. They are normalized value between 0 and 1. \( \gamma (0 \leq \gamma \leq 1) \) is weight of two values. If \( \gamma = 0.5 \), the equation is harmonic mean of two values. Therefore \( I(x_k) \) becomes large when two values are equally large.

3 Experiments and Evaluation

In this section, we describe our experiments and evaluate proposed methods.

We select a subtree of MeSH thesaurus for the experiment. “metabolic diseases(C18.452)” node is root of the subtree, and the subtree consists of 436 disease names which are target terms for specificity measuring. We used MEDLINE database corpus (170,000 abstracts, 20,000,000 words) to extract statistical information.

Each method was evaluated by two criteria, coverage and precision. Coverage is the fraction of the terms which have the specificity by given method. Method 2 gets relatively lower coverage than method 1, because method 2 can measure the specificity only when both the terms and their modifiers occur in corpus. Method 1 can measure the specificity whenever parts of composite words appear in corpus. Precision is the fraction of correct specificity relations values as equation (13).

\[
p = \frac{\# \text{ of } R(p,c) \text{ with correct specificity}}{\# \text{ of all } R(p,c)}
\]

(13)

where \( R(p,c) \) is a parent-child relation in MeSH thesaurus. If child term \( c \) has larger specificity than that of parent term \( p \), then the relation is said to have correct specificity. We divided parent-child relations into two types. Relations where parent term is nested in child term are categorized as type I. Other relations are categorized as type II. There are 43 relations in type I and 393 relations in type II. The relations in type I always have correct specificity provided modifier-head information method described in section 2.1 is applied.

We tested prior experiment for 10 human subjects to find out the upper bound of precision. The subjects are all medical doctors of internal medicine, which is closely related division to “metabolic diseases”. They were asked to identify parent-child relationship for given term pairs. The average precisions of type I and type II were 96.6% and 86.4% respectively. We set these values as upper bound of precision for suggested methods.

\[\text{ICmp(x_k)}\]

\[\text{ICtx(x_k)}\]

\[\text{MEDLINE}\]

\[\text{MEDLINE is a database of biomedical articles serviced by National Library of Medicine, USA. (http://www.nlm.nih.gov)}\]
The specificity of terms was measured with method 1, method 2, and method 3 as Table 2. Two additional methods, based on term frequency and term tf.idf, were experimented to compare compositionality based methods and term based methods.

Method 1 showed better performance than term based methods. This result illustrate 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. Word tf.idf based method showed better precision than word frequency based method. This result illustrate that tf.idf of words is more informative than frequency of words.

Method 3 showed the best precision, 82.0%, because the two methods interacted complementary. In hybrid method, the weight value $\gamma = 0.8$ indicates that compositional information is more informative than contextual information for the specificity of domain specific terms.

One reason of the errors is that the names of some internal nodes in MeSH thesaurus are category names rather disease names. For example, as “acid-base imbalance (C18.452.076)” is name of disease category, it doesn't occur as frequently as other real disease names. Other predictable reason is that we didn’t consider various surface forms of same term. For example, although “NIDDM” is acronym of “non insulin dependent diabetes mellitus”, the system counted two terms separately. Therefore the extracted statistics can’t properly reflect semantic level information.

### Table 2. Experimental results (%)

| Methods                                      | Precision | Coverage |
|----------------------------------------------|-----------|----------|
| Human subjects (Average)                     |           |          |
| Term frequency                               | 96.6      | 86.4     | 87.4     |
| Term tf.idf                                  | 100.0     | 53.5     | 60.6     | 89.5     |
| Word Freq.                                   | 37.2      | 72.5     | 69.0     | 100.0    |
| Word Freq. + Structure ($\alpha$=$\beta$=0.2) | 100.0     | 72.8     | 75.5     | 100.0    |
| Word tf.idf + Structure ($\alpha$=$\beta$=0.2) | 44.2      | 75.3     | 72.2     | 100.0    |
| Contextual Information Method (Method 2) (mod cnt>1) | 90.0      | 66.4     | 70.0     | 70.2     |
| Hybrid Method (Method 3) (tf.idf + Struct, $\gamma$=0.8) | 95.0      | 79.6     | 82.0     | 70.2     |

As the proposed methods don’t use domain dependent information, they can be adapted to other domains without extra processes.

In the future, we will modify the system to handle various term formations such as abbreviated form. Finally we will apply the proposed methods to the terms of other specific domains.

### 5 Acknowledgements

This work was supported in part by Ministry of Science & Technology, Ministry of Culture & Tourism of Korean government, and Korea Science & Engineering Foundation.

### References

Caraballo, S. A., Charniak, E. 1999. Determining the Specificity of Nouns from Text. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora

Conexor. 2004. Conexor Functional Dependency Grammar Parser. http://www.conexor.com

Croft, W. 2004. Typology and Universals. 2 nd ed. Cambridge Textbooks in Linguistics, Cambridge Univ. Press

Franz, K., et al. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. Journal of Digital Libraries, vol. 3, num. 2

Grefenstette, G. 1994. Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers

Haykin, S. 1994. Neural Network. IEEE Press

Manning, C. D. and Schutze, H. 1999. Foundations of Statistical Natural Language Processing. The MIT Press

Pereira, F., Tishby, N., and Lee, L. 1993. Distributational clustering of English words. Proceedings of ACL

Sanderson, M. 1999. Deriving concept hierarchies from text. Proceedings of ACM SIGIR