Determining the Specificity of Terms using Compositional and Contextual Information

Pum-Mo Ryu
Department of Electronic Engineering and Computer Science
KAIST
Pum-Mo.Ryu@kaist.ac.kr

Abstract

This paper introduces new specificity determining methods for terms using compositional and contextual information. Specificity of terms is the quantity of domain specific information that is contained in the terms. The methods are modeled as information theory like measures. As the methods don't use domain specific information, they can be applied to other domains without extra processes. Experiments showed very promising result with the precision of 82.0% when the methods were applied to the terms in MeSH thesaurus.

1. Introduction

Terminology management concerns primarily with *terms*, i.e., the words that are assigned to concepts used in domain-related texts. A *term* is a meaningful unit that represents a specific concept within a domain (Wright, 1997).

Specificity of a term represents the quantity of domain specific information contained in the term. If a term has large quantity of domain specific information, specificity value of the term is large; otherwise specificity value of the term is small. Specificity of term X is quantified to positive real number as equation (1).

$$Spec(X) \in R^{+} \quad (1)$$

Specificity of terms is an important necessary condition in term hierarchy, *i.e.*, if X_1 is one of ancestors of X_2, then $Spec(X_1)$ is less than $Spec(X_2)$. Specificity can be applied in automatic construction and evaluation of term hierarchy.

When domain specific concepts are represented as 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 their meaning. In the second category, new terms are created independently to 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 used to discriminate the features 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] tree. Node numbers represent hierarchical structure of terms

Contextual information has been mainly used to represent the characteristics of terms. (Caraballo, 1999A) (Grefenstette, 1994) (Hearst, 1992) (Pereira, 1993) and (Sanderson, 1999) used contextual information to find hyponymy relation between terms. (Caraballo, 1999B) also used contextual information to determine the specificity of nouns. Contrary, compositional information of terms has not been commonly discussed.

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

We propose new specificity measuring methods based on both compositional and contextual information. The methods are formulated as information theory like measures. Because the methods don't use domain specific information, they are easily adapted to terms of other domains.

This paper consists as follow: compositional and contextual information is discussed in section 2, information theory like measures are described in section 3, experiment and evaluation is discussed in section 4, finally conclusions are drawn in section 5.

2. Information for Term Specificity

In this section, we describe compositional information and contextual information.

2.1. Compositional Information

By compositionality, the meaning of whole term can be strictly predicted from the meaning of the individual words (Manning, 1999). Many terms are created by appending modifiers to existing terms. In this mechanism, features of modifiers are added to features of existing terms to make new concepts. Word frequency and tf.idf value are used to quantify features of unit words. Internal modifier-head structure of terms is used to measure specificity incrementally.

We assume that terms composed of low frequency words have large quantity of domain information. Because low frequency words appear only in limited number of terms, the words can clearly discriminate the terms to other terms.

tf.idf, multiplied value of term frequency (tf) and inverse document frequency (idf), is widely used term weighting scheme in information retrieval (Manning, 1999). Words with high term frequency and low document frequency get large tf.idf value. Because a document usually discusses one topic, and words of large tf.idf values are good index terms for the document, the words are considered to have topic specific information. Therefore, if a term includes words of large tf.idf value, the term is assumed to have topic or domain specific information.

If the modifier-head structure of a term is known, the specificity of the term is calculated incrementally starting from head noun. In this manner, specificity value of a term is always larger than that of the base (head) term. This result answers to the assumption that more specific term has larger specificity value. However, it is very difficult to analyze modifier-head structure of compound noun. We use simple nesting relations between terms to analyze structure of terms. A term X is nested to term Y, when X is substring of Y (Frantzi, 2000) as follows:

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

For example two terms, "*diabetes mellitus*" and "*insulin dependent diabetes mellitus*", are all disease names, and the former is nested in the latter. In this case, "*diabetes mellitus*" is base term and "*insulin dependent*" is modifier of "*insulin dependent diabetes mellitus*" by definition 1. If multiple terms are nested in a term, the longest term is selected as head term. Specificity of Y is measured as equation (2).

$$Spec(Y) = Spec(X) + \alpha \cdot Spec(W_1) + \beta \cdot Spec(W_2) \quad (2)$$

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

2.2. Contextual Information

There are some problems that are hard to address using compositional information alone. Firstly, although features of "*wolfram syndrome*" share many common features with features of "*insulin-dependent diabetes mellitus*" in semantic level, they don't share any common words in lexical level. In this case, it is unreasonable to compare two specificity values measured based on compositional information alone. Secondly, when several words are combined to a 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 "*diabetes mellitus*" from two separate words "*wolfram*" and "*syndrome*". Finally, modifier-head structure of some terms is ambiguous. For instance, "*vampire slayer*" might be a slayer who is vampire or a slayer of vampires. Therefore contextual is used to complement these problems.

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

General terms usually tend to be modified by other words. Contrary, domain specific terms don't tend to be modified by other words, because they have sufficient information in themselves (Caraballo, 1999B). Under this assumption, we use probabilistic distribution of modifiers as contextual information. Because domain specific terms, unlike general words, are rarely modified in corpus, it is important to collect statistically sufficient modifiers from given corpus. Therefore accurate text processing, such as syntactic parsing, is needed to extract modifiers. As Caraballo's work was for general words, they extracted only rightmost prenominals as context information. We use Conexor functional dependency parser (Conexor, 2004) to analyze the structure of sentences. Among many dependency functions defined in Conexor parser, "*attr*" and "*mod*" functions are used to extract modifiers from analyzed structures. If a term or modifiers of the term do not occur in corpus, specificity of the term can not be measured using contextual information

3. Specificity Measuring Methods

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

In information theory, when a message with low probability occurs on channel output, the amount of *surprise* is large, and the length of bits to represent this message becomes long. Therefore the large quantity of information is gained by this message (Haykin, 1994). 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 (3) for further explanation.

$$T = \{t_k \mid 1 \le k \le n\} \tag{3}$$

where t_k is a term and n is total number of terms. In next step, a discrete random variable X is defined as equation (4).

$$\begin{aligned} X &= \{x_k \mid 1 \le k \le n\} \\ p(x_k) &= \mathrm{Prob}(X = x_k) \end{aligned} \tag{4}$$

where x_k is an event of a term t_k occurs in corpus, $p(x_k)$ is the probability of event x_k. The information quantity, $I(x_k)$, gained after observing the event x_k, is defined by the logarithmic function. Finally $I(x_k)$ is used as specificity value of t_k as equation (5).

$$Spec(t_k) \approx I(x_k) = -\log p(x_k) \tag{5}$$

In equation (5), we can measure specificity of t_k, by estimating $p(x_k)$. We describe three estimating methods of $p(x_k)$ in following sections.

3.1. Compositional Information based Method (Method 1)

In this section, we describe a method using compositional information introduced in section 2.1. 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 detail description, we assume that a term t_k consists of one or more words as equation (6).

$$t_k = w_1 w_2 ... w_m \tag{6}$$

where w_i is *i-th* word in t_k. In next step, a discrete random variable Y is defined as equation (7).

$$\begin{aligned} Y &= \{y_i \mid 1 \le i \le m\} \\ p(y_i) &= \mathrm{Prob}(Y = y_i) \end{aligned} \tag{7}$$

where y_i is an event of a word w_i occurs in term t_k, $p(y_i)$ is the probability of event y_i. Information quantity, $I(x_k)$, in equation (5) is redefined as equation (8) based on previous assumption.

$$I(x_k) = -\sum_{i=1}^{m} p(y_i) \log p(y_i) \tag{8}$$

where $I(x_k)$ is average information quantity of all words in t_k. Two information sources, word frequency, tf.idf are used to estimate $p(y_i)$. In this

mechanism, $p(y_i)$ for informative words should be smaller than that of non informative words.

When word frequency is used to quantify features of words, $p(y_i)$ in equation (8) is estimated as equation (9).

$$p(y_i) \approx p_{MLE}(w_i) = \frac{freq(w_i)}{\sum_j freq(w_j)} \quad (9)$$

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

When tf.idf is used to quantify features of words, $p(y_i)$ in equation (8) is estimated as equation (10).

$$p(y_i) \approx p_{MLE}(w_i) = 1 - \frac{tf \cdot idf(w_i)}{\sum_j tf \cdot idf(w_j)} \quad (10)$$

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

3.2. Contextual Information based Method (Method 2)

In this section, we describe a method using contextual information introduced in section 2.2. Entropy of probabilistic distribution of modifiers for a term is defined as equation (11).

$$H_{mod}(t_k) = -\sum_i p(mod_i, t_k) \log p(mod_i, t_k) \quad (11)$$

where $p(mod_i,t_k)$ is the probability of mod_i modifies t_k and is estimated as equation (12).

$$p_{MLE}(mod_i, t_k) = \frac{freq(mod_i, t_k)}{\sum_j freq(mod_j, t_k)} \quad (12)$$

where $freq(mod_i,t_k)$ is number of frequencies that mod_i modifies t_k in corpus, j is index of all modifiers of t_k in corpus. The entropy calculated by equation (11) is the average information quantity of all (mod_i,t_k) pairs. Specific terms have low entropy, because their modifier distributions are simple. Therefore inversed entropy is assigned to $I(x_k)$ in equation (5) to make specific terms get large quantity of information as equation (13).

$$I(x_k) \approx \max_{1 \le i \le n}(H_{mod}(t_i)) - H_{mod}(t_k) \quad (13)$$

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

3.3. Hybrid Method (Method 3)

In this section, we describe a hybrid method to overcome shortcomings of previous two methods. This method measures term specificity as equation (14).

$$I(x_k) \approx \frac{1}{\gamma(\frac{1}{I_{Cmp}(x_k)}) + (1-\gamma)(\frac{1}{I_{Ctx}(x_k)})} \quad (14)$$

where $I_{Cmp}(x_k)$ and $I_{Ctx}(x_k)$ are normalized $I(x_k)$ values between 0 and 1, which are measured by compositional and contextual information based methods respectively. $\gamma(0 \le \gamma \le 1)$ 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.

4. Experiment and Evaluation

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

4.1. Evaluation

A sub-tree of MeSH thesaurus is selected for experiment. "*metabolic diseases(C18.452)*" node is 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 MEDLINE[2] database using the disease names as quires. Therefore, all the abstracts are related to some of the disease names. The set consists of about 170,000 abstracts (20,000,000 words). The abstracts are analyzed using 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.

The system was evaluated by two criteria, coverage and precision. Coverage is the fraction

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

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 Freq.	0.37	72.5	69.0	100.0
	Word Freq.+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
Contextual Information Method (Method 2) (mod cnt>1)		90.0	66.4	70.0	70.2
Hybrid Method (Method 3) (tf·idf + Struct, γ=0.8)		95.0	79.6	82.0	70.2

Table 2. Experimental results (%)

of the terms which have specificity values by given measuring method as equation (15).

$$c = \frac{\#\ of\ terms\ with\ specificity}{\#\ of\ all\ terms} \quad (15)$$

Method 2 gets relatively lower coverage than method 1, because method 2 can measure specificity when both the terms and their modifiers appear in corpus. Contrary, method 1 can measure specificity of the terms, when parts of unit words appear in corpus. Precision is the fraction of relations with correct specificity values as equation (16).

$$p = \frac{\#\ of\ R(p,c)\ with\ correct\ specificity}{\#\ of\ all\ R(p,c)} \quad (16)$$

where *R(p,c)* is a parent-child relation in MeSH thesaurus, and this relation is valid only when specificity of two terms are measured by given method. If child term *c* has larger specificity value than that of parent term *p*, then the relation is said to have correct specificity values. 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 values provided structural information method described 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 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 upper bound of precision for suggested methods.

Specificity values of terms were measured with method 1, method 2, and method 3 as Table 2. In method 1, word frequency based method, word tf.idf 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 compositionality based method and whole term based method. Two methods which showed the best performance in method 1 and method 2 were combined into method 3.

Word frequency and tf.idf based method showed better performance than term based methods. This result indicates that the information of terms is divided into unit words rather than into whole terms. This result also 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 2 showed the best performance, precision 70.0% and coverage 70.2%, when we counted modifiers which modify the target terms two or more times. However, method 2 showed worse performance than word tf.idf and structure based method. It is assumed that sufficient contextual information for terms was not collected from corpus, because domain specific terms are rarely modified by other words.

Method 3, hybrid method of method 1 (tf.idf of words, structure information) and method 2, showed the best precision of 82.0% of all, because the two methods interacted complementary.

The coverage of this method was 70.2% which equals to the coverage of method 2, because the specificity value is measured only when the specificity of method 2 is valid. In hybrid method, the weight value $\gamma = 0.8$ indicates that compositional information is more informatives than contextual information when measuring the specificity of domain-specific terms. The precision of 82.0% is good performance compared to upper bound of 87.4%.

4.2. Error Analysis

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 independently. Therefore the extracted statistics can't properly reflect semantic level information.

If we analyze morphological structure of terms, some errors can be reduced by internal structure method described in section 2.1. For example, "*nephrocalcinosis*" have modifier-head structure in morpheme level; "*nephro*" is modifier and "*calcinosis*" is head. Because word formation rules are heavily dependent on the domain specific morphemes, additional information is needed to apply this approach to other domains.

5. Conclusions

This paper proposed specificity measuring methods for terms based on information theory like measures using compositional and contextual information of terms. The methods are experimented on the terms in MeSH thesaurus. Hybrid method showed the best precision of 82.0%, because two methods complemented each other. As the proposed methods don't use domain dependent information, the methods easily can be adapted to other domains.

In the future, the system will be modified to handle various term formations such as abbreviated form. Morphological structure analysis of words is also needed to use the morpheme level information. Finally we will apply the proposed methods to terms of other domains and terms in general domains such as WordNet.

Acknowledgements

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

References

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

Caraballo, S. A. and Charniak, E. 1999B. *Determining the Specificity of Nouns from Text*. In the 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

Frantzi, K., Anahiadou, S. and Mima, H. 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, pp. 444

Hearst, M. A. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*. In proceedings of ACL

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

Pereira, F., Tishby, N., and Lee, L. 1993. *Distributational clustering of English words*. In the proceedings of ACL

Sanderson, M. 1999. *Deriving concept hierarchies from text*. In the Proceedings of the 22th Annual ACM S1GIR Conference on Research and Development in Information Retrieval

Wright, S. E., Budin, G.. 1997. *Handbook of Term Management: vol. 1*. John Benjamins publishing company

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