

# Taxonomy Learning using Term Specificity and Similarity

**Pum-Mo Ryu**

Computer Science Division, KAIST  
KORTERM/BOLA  
Korea  
pmryu@world.kaist.ac.kr

**Key-Sun Choi**

Computer Science Division, KAIST  
KORTERM/BOLA  
Korea  
kschoi@cs.kaist.ac.kr

## Abstract

Learning taxonomy for technical terms is difficult and tedious task, especially when new terms should be included. The goal of this paper is to assign taxonomic relations among technical terms. We propose new approach to the problem that relies on term specificity and similarity measures. Term specificity and similarity are necessary conditions for taxonomy learning, because highly specific terms tend to locate in deep levels and semantically similar terms are close to each other in taxonomy. We analyzed various features used in previous researches in view of term specificity and similarity, and applied optimal features for term specificity and similarity to our method.

## 1 Introduction

Taxonomy is a collection of controlled vocabulary terms organized into a hierarchical structure. Each term in a taxonomy is one or more parent-child relationships to other terms in the taxonomy. Taxonomies are useful artifacts for organizing many aspects of knowledge. As components of ontologies, taxonomies can provide an organizational model for a domain (domain ontology), or a model suitable for specific tasks (task ontologies) (Burgun & Bodenreider, 2001). However their wide usage is still hindered by time-consuming, cost-ineffective building processes.

The main paradigms of taxonomy learning are on the one hand pattern based approaches and on the other hand distributional hypothesis based approaches. The former is approaches based on matching lexico-syntactic patterns which convey

taxonomic relations in a corpus (Hearst, 1992; Iwanska et al., 2000), and the latter is statistical approaches based on the distribution of context in corpus (Cimiano et al., 2005; Yamamoto et al., 2005; Sanderson & Croft, 1999). The former features a high precision and low recall compared to the latter. The quality of learned relations is higher than those of statistical approaches, while the patterns are rarely applied in real corpus. It is also difficult to improve performance of pattern based approaches because they are simple and clear. So, many researches have been focused on raising precision of statistical approaches.

We introduce new distributional hypothesis based taxonomy learning method using term specificity and term similarity. Term specificity is a measure of information quantity of terms in given domain. When a term has much domain information, the term is highly specific to the domain, and vice versa (Ryu & Choi, 2005). Because highly specific terms tend to locate in low level in domain taxonomy, term specificity can be used as a necessary condition for taxonomy learning. Term similarity is degree of semantic overlap among terms. When two terms share many common characteristics, they are semantically similar to each other. Term similarity can be another necessary condition for taxonomy learning, because semantically similar terms locate near by in given domain taxonomy. The two conditions are generally valid for terms in a taxonomic relation, while terms satisfying the conditions do not always have taxonomic relation. So they are necessary conditions for taxonomy learning.

Based on these conditions, it is highly probable that term  $t_1$  is an ancestor of term  $t_2$  in domain taxonomy  $T_D$ , when  $t_1$  and  $t_2$  are semantically similar enough and the specificity of  $t_1$  is lower than that of  $t_2$  in  $D$  as in Figure 1. However,  $t_1$  is not an ancestor of  $t_3$  even though the speci-

ficiency of  $t_1$  is lower than that of  $t_3$  because  $t_1$  is not similar to  $t_3$  on the semantic level.

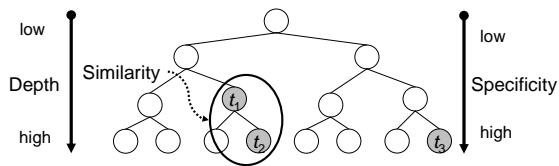


Figure 1. Term specificity and term similarity in a domain taxonomy  $T_D$

The strength of this method lies in its ability to adopt different optimal features for term specificity and term similarity. Most of current researches relied on single feature such as adjectives of terms, verb-argument relation, or co-occurrence ratio in documents according to their methods. Firstly, we analyze characteristics of features for taxonomy learning in view of term specificity and term similarity to show that the features embed characteristics of specificity and similarity, and finally apply optimal features to our method.

Additionally we tested inside information of terms to measure term specificity and similarity. As multiword terms cover the larger part of technical terms, lexical components are featuring information representing semantics of terms (Cerbah, 2000).

The remainder of this paper is organized follows. Characteristics of term specificity are described in Section 2, while term similarity and its features are addressed in Section 3. Our taxonomy learning method is discussed in Section 4. Experiment and evaluation are discussed in Section 5, and finally, conclusions are drawn in Section 6.

## 2 Term Specificity

Specificity is degree of detailed information of an object about given target object. For example, if an encyclopedia contains detailed information about '*IT domain*', then the encyclopedia is '*IT specific encyclopedia*'. In this context, specificity is a function of objects and target object to real number. Traditionally term specificity is widely used in information retrieval systems to weight index terms in documents (S. Jones, 1972; Aizawa, 2003; Wong & Yao, 1992). In information retrieval context, term specificity is function of index terms and documents. On the other hand, term specificity is the function of terms and target domains in taxonomy learning context (Ryu & Choi 2005). Term specificity to a domain is

quantified to a positive real number as shown in Eq. (1).

$$Spec(t | D) \in R^+ \quad (1)$$

where  $t$  is a term, and  $Spec(t|D)$  is the specificity of  $t$  in a given domain  $D$ . We simply use  $Spec(t)$  instead of  $Spec(t|D)$  assuming a particular domain  $D$  in this paper.

Understanding the relation between domain concepts and their lexicalization methods is needed, before we describe term specificity measuring methods. Domain specific concepts can be distinguished by a set of what we call '*characteristics*'. More specific concepts are created by adding characteristics to the set of characteristics of existing concepts. Let us consider two concepts:  $C_1$  and  $C_2$ .  $C_1$  is an existing concept and  $C_2$  is a newly created concept by combining new characteristics to the characteristic set of  $C_1$ . In this case,  $C_1$  is an ancestor of  $C_2$  (ISO, 2000). When domain specific concepts are lexicalized as terms, the terms' word-formation is classified into two categories based on the composition of component words. In the first category, new terms are created by adding modifiers to existing terms. Figure 2 shows a subtree of financial ontology. For example '*current asset*' was created by adding the modifier '*current*' to its hypernym '*asset*'. In this case, inside information is a good evidence to represent the characteristics. In the second category, new terms are created independently of existing terms. For example, '*cache*', '*inventory*', and '*receivable*' share no common words with their hypernyms '*current asset*' and '*asset*'. In this case, outside information is used to differentiate the characteristics of the terms.

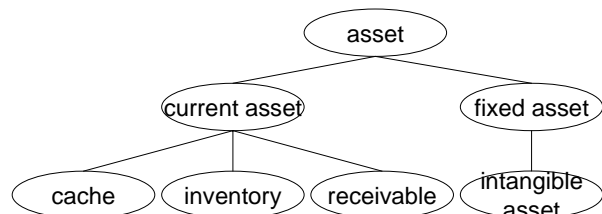


Figure 2. Subtree of financial ontology

There are many kinds of inside and outside information to be used in measuring term specificity. Distribution of adjective-term relation and verb-argument dependency relation are collocation based statistics. Distribution of adjective-term relation refers to the idea that specific nouns are rarely modified, while general nouns are fre-

quently modified in text. This feature has been discussed to measure specificity of nouns in (Caraballo, 1999; Ryu & Choi, 2005) and to build taxonomy of Japanese nouns (Yamamoto et al., 2005). Inversed specificity of a term can be measured by entropy of adjectives as shown Eq. (2).

$$Spec_{adj}(t)^{-1} = -\sum_{adj} P(adj|t) \log P(adj|t) \quad (2)$$

where  $P(adj|t)$ , the probability that  $adj$  modifies  $t$ , is estimated as  $freq(adj,t)/freq(t)$ . The entropy is the average information quantity of all  $(adj,t)$  pairs for term  $t$ . Specific terms have low entropy, because their adjective distributions are simple.

For verb-argument distribution, we assume that domain specific terms co-occur with selected verbs which represent special characteristics of terms while general terms are associated with multiple verbs. Under this assumption, we make use of syntactic dependencies between verbs appearing in the corpus and their arguments such as subjects and objects. For example, ‘*inventory*’<sup>1</sup>, in Figure 2, shows a tendency to be objects of specific verbs like ‘*increase*’ and ‘*reduce*’. This feature was used in (Cimiano et al., 2005) to learn concept hierarchy. Inversed specificity of a term can be measured by entropy of verb-argument relations as Eq. (3).

$$Spec_{v_{arg}}(t)^{-1} = -\sum_{v_{arg}} P(t|v_{arg}) \log P(t|v_{arg}) \quad (3)$$

where  $P(t|v_{arg})$ , the probability that  $t$  is argument of  $v_{arg}$ , is estimated as  $freq(t,v_{arg})/freq(v_{arg})$ . The entropy is the average information quantity of all  $(t,v_{arg})$  pairs for term  $t$ .

Conditional probability of term co-occurrence in documents was used in (Sanderson & Croft, 1999) to build term taxonomy. This statistics is based on the assumption that, for two terms,  $t_i$  and  $t_j$ ,  $t_i$  is said to subsume  $t_j$  if the following two conditions hold,

$$P(t_i|t_j) = 1 \text{ and } P(t_j|t_i) < 1 \quad (4)$$

In other words,  $t_i$  subsumes  $t_j$  if the documents which  $t_j$  occurs in are a subset of the documents which  $t_i$  occurs in, therefore  $t_i$  can be parent of  $t_j$  in taxonomy. Although a good number of term pairs are found that adhere to the two subsump-

tion conditions, it is noticed that many are just failing to be included because a few occurrences of the subsumed term,  $t_j$ , does not co-occur with  $t_i$ . Subsequently, the conditions are relaxed and *subsume* function is defined as Eq. (5). In case of  $P(t_i|t_j) > P(t_j|t_i)$ , *subsume*( $t_i, t_j$ ) returns 1, otherwise returns 0.

$$subsume(t_i, t_j) = \begin{cases} 1 & \text{if } P(t_i|t_j) > P(t_j|t_i) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

We apply this function to calculate term specificity as shown Eq. (6) where a term is specific when it is subsumed by most of other terms. Specificity of  $t$  is determined by the ratio of terms that subsume  $t$  over all co-occurring terms.

$$Spec_{coldoc}(t) = \frac{\sum_{1 \leq j \leq n} subsume(t_j, t)}{n} \quad (6)$$

where  $n$  is number of terms co-occurring terms with  $t$ .

Finally, inside-word information is important to compute specificity for multiword terms. Consider a term  $t$  that consists of two words like  $t = w_1 w_2$ . Two words,  $w_1$  and  $w_2$ , have their unique characteristics and the characteristics are summed up to the characteristic of  $t$ . Mutual information is used to estimate the association between a term and its component words. Let  $T = \{t_1, \dots, t_N\}$  be a set of terms found in a corpus, and  $W = \{w_1, \dots, w_M\}$  be a set of component words composing the terms in  $T$ . Assume a joint probability distribution  $P(t_i, w_j)$ , probability of  $w_j$  is a component of  $t_i$ , is given for  $t_i$  and  $w_j$ . Mutual information between  $t_i$  and  $w_j$  compares the probability of observing  $t_i$  and  $w_j$  together and the probability of observing  $t_i$  and  $w_j$  independently. The mutual information represents the reduction of uncertainty about  $t_i$  when  $w_j$  is observed. The summed mutual information between  $t_i$  and  $W$ , as in Eq. (7), is total reduction of uncertainty about  $t_i$  when all component words are observed.

$$Spec_{in}(t_i) = \sum_{w_j \in W} \log \frac{P(t_i, w_j)}{P(t_i)P(w_j)} \quad (7)$$

This equation indicates that  $w_j$  which is highly associated to  $t_i$  contributes specificity of  $t_i$ . For example, ‘*debenture bond*’ is more specific concept than ‘*financial product*’. Intuitively, ‘*debenture*’ is highly associated to ‘*debenture bond*’

<sup>1</sup> ‘*Inventory*’ consists of a list of goods and materials held available in stock (<http://en.wikipedia.org/wiki/Inventory>).

compared with ‘bond’ to ‘debenture bond’ or ‘financial’, ‘product’ to ‘financial product’.

### 3 Term Similarity

We evaluate four statistical and lexical features, related to taxonomy learning, in view of term similarity. Three statistical features have been used in existing taxonomy learning researches.

(Sanderson & Croft, 1999) used conditional probability of co-occurring terms in same document in taxonomy learning process as shown in Eq. (4). This feature can be used to measure similarity of terms. If two terms co-occur in common documents, they are semantically similar to each other. Based on this assumption, we can calculate term similarity by comparing the frequency of co-occurring  $t_i$  and  $t_j$  together and the frequency of occurring  $t_i$  and  $t_j$  independently, as Eq. (8).

$$Sim_{coldoc}(t_i, t_j) = \frac{2 * df(t_i, t_j)}{df(t_i) + df(t_j)} \quad (8)$$

where  $df(t_i, t_j)$  is number of documents in which both  $t_i$  and  $t_j$  co-occur,  $df(t_i)$  is number of documents in which  $t_i$  occurs.

(Yamamoto et al., 2005) used adjective patterns to make characteristics vectors for terms in Complementary Similarity Measure (CSM). Although CSM was initially designed to extract superordinate-subordinate relations, it is a similarity measure by itself. They proposed two CSM measures; one is for binary images in which values in feature vectors are 0 or 1, and the other is for gray-scale images in which values in feature vectors are 0 through 1. We adapt gray-scale measure in similarity calculation, because it showed better performance in their research.

(Cimiano et al., 2005) applied Formal Concept Analysis (FCA) to extract taxonomies from a text corpus. They modeled the context of a term as a vector representing syntactic dependencies. Similarity based on verb-argument dependencies is calculated using cosine measure as Eq. (9).

$$Sim_{varg}(t_i, t_j) = \frac{\sum_{v_{arg} \in V} P(t_i | v_{arg}) P(t_j | v_{arg})}{\sqrt{\sum_{v_{arg} \in V} P(t_i | v_{arg})^2} \sqrt{\sum_{v_{arg} \in V} P(t_j | v_{arg})^2}} \quad (9)$$

where  $P(t|v_{arg})$ , the probability that  $t$  is argument of  $v_{arg}$ , is estimated as  $freq(t, v_{arg})/freq(v_{arg})$ . Above three similarity measures are valid when

terms,  $t_i$  and  $t_j$ , appear in corpus one or more times.

The last similarity measure is based on inside information of terms. Because many domain terms are multiword terms, component words are clues for term similarity. If two terms share many common words, they share common characteristics in given domain. For example, four words ‘asset’, ‘current asset’, ‘fixed asset’ and ‘intangible asset’ share characteristics related to ‘asset’ as in Figure 2. This similarity measure is shown in Eq. (10).

$$Sim_{in}(t_i, t_j) = \frac{2 * cwc(t_i, t_j)}{|t_i| + |t_j|} \quad (10)$$

where  $|t|$  is word count of  $t$ , and  $cwc(t_i, t_j)$  is common word count in  $t_i$  and  $t_j$ .  $Sim_{in}(t_i, t_j)$  is valid when  $cwc(t_i, t_j) > 0$ . Because  $cwc(t_i, t_j) = 0$  for most of term pairs, it is difficult to catch reliable results for all possible term pairs.

### 4 Taxonomy Learning Process

We model taxonomy learning process as a sequential insertion of new terms to current taxonomy. New taxonomy starts with empty state, and changes to rich taxonomic structure with the repeated insertion of terms as depicted in Figure 3. Terms to be inserted are sorted by term specificity values. Term insertion based on the increasing order of term specificity is natural, because the taxonomy grows from top to down with term insertion process in increasing specificity sequence.

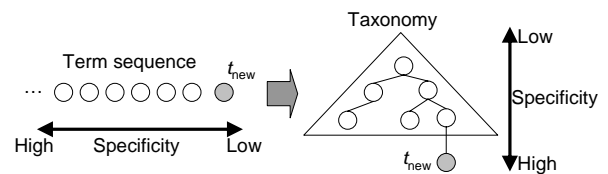


Figure 3. Terms are inserted to taxonomy in the sequence of specificity

According to above assumption, our system selects possible hypernyms of a new term,  $t_{new}$  in current taxonomy as following steps:

- Step 1: Select  $n$ -most similar terms to  $t_{new}$  from current taxonomy
- Step 2: Select candidate hypernyms of  $t_{new}$  from  $n$ -most similar terms. Specificity of candidate hypernyms is less than that of  $t_{new}$ .

- Step 3: Insert  $t_{new}$  as hyponyms of candidate hypernyms

For example, suppose  $t_2$ ,  $t_4$ ,  $t_5$  and  $t_6$ , are four most similar terms to  $t_{new}$  in Figure 4. Two terms  $t_2$  and  $t_4$  are selected as candidate hypernyms of  $t_{new}$ , because specificity of the terms is less than specificity of  $t_{new}$ .

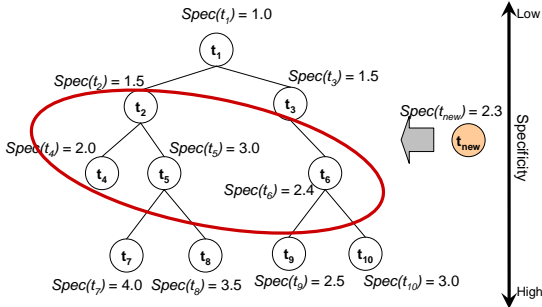


Figure 4. Selection of candidate hypernyms of  $t_{new}$  from taxonomy using term specificity and similarity

## 5 Experiment and Evaluation

We applied our taxonomy learning method to set of terms in existing taxonomy. We removed all relations from the taxonomy, and made new taxonomic relations among the terms. The learned taxonomy was then compared to original taxonomy. Our experiment is composed of four steps. Firstly, we calculated term specificity using specificity measures discussed in chapter 2, secondly, we calculated term similarity using similarity measures described in chapter 3, thirdly, we applied the best specificity and similarity features to our taxonomy building process, and finally, we evaluated our method and compared with other taxonomy learning methods.

Finance ontology<sup>2</sup> which was developed within the GETESS project (Staab et al., 1999) was used in our experiment. We slightly modified original ontology. We unified different expressions of same concept to identical expression. For example, 'cd-rom drive' and 'cdrom drive' are unified as 'cd-rom drive' because the former is more usual expression than the latter. We also removed terms that are not descendents of 'root' node to make the taxonomy have single root node. The taxonomy consists of total 1,819 nodes and 1,130 distinct nodes. Maximum and average depths are 15 and 5.5 respectively, and

maximum and average children nodes are 32 and 3.5 respectively.

We considered Reuters21578<sup>3</sup> corpus, over 3.1 million words in title and body fields. We parsed the corpus using Connexor functional dependency parser<sup>4</sup> and extracted various statistics: term frequency, distribution of adjectives, distribution of co-occurring frequency in documents, and verb-argument distribution.

### 5.1 Term Specificity

Term specificity was evaluated based on three criteria: recall, precision and F-measure. Recall is the fraction of the terms that have specificity values by the given measuring method. Precision is the fraction of relations with correct specificity values. F-measure is a harmonic mean of precision and recall into a single measure of overall performance. Precision ( $P_{spec}$ ), recall ( $R_{spec}$ ), F-measure ( $F_{spec}$ ) is defined as follows:

$$R_{spec} = \frac{\# \text{ of terms with specificity}}{\# \text{ of all terms}} \quad (11)$$

$$P_{spec} = \frac{\# \text{ of } R_{valid}(p,c) \text{ with correct specificity}}{\# \text{ of } R_{valid}(p,c)}$$

where  $R_{valid}(p,c)$  is a valid parent-child relation in original taxonomy, and a relation is *valid* when the specificity of two terms are measured by the given method. If the specificity of child term,  $c$ , is larger than that of parent term,  $p$ , then the relation is *correct*.

We tested four specificity measuring methods discussed in section 2 and the result is shown in Table 1.  $Spec_{adj}$  showed the highest precision as we anticipated. Because domain specific terms have sufficient information in themselves; they are rarely modified by other words in real text. However,  $Spec_{adj}$  showed the lowest recall for data sparseness problem. As mentioned above, it is hard to collect sufficient adjectives for domain specific terms from text.  $Spec_{varg}$  showed the lowest precision. This result indicates that distribution of verb-argument relation is less correlated to term specificity.  $Spec_{in}$  showed the highest recall because it measures term specificity using component words contrary to other methods.  $Spec_{coldoc}$  showed comparable precision and recall.

<sup>2</sup> The ontology can be downloaded at <http://www.aifb.uni-karlsruhe.de/WBS/pci/FinanceGoldStandard.isa>. P. Cimiano and his colleagues added English labels for the originally German labeled nodes (Cimiano et al., 2005)

<sup>3</sup> <http://www.daviddlewis.com/resources/testcollections/reuters21578/>

<sup>4</sup> <http://www.connexor.com/>

We harmonized  $Spec_{in}$  and  $Spec_{adj}$  to  $Spec_{in/adj}$  as described in (Ryu & Choi, 2005) to take advantages of both inside and outside information. Harmonic mean of two specificity values was used in  $Spec_{in/adj}$  method.  $Spec_{in/adj}$  showed the highest F-measure because precision was higher than that of  $Spec_{in}$  and recall was equal to that of  $Spec_{in}$ .

Table 1. Precision, recall and F-measure for term specificity

Method	Precision	Recall	F-measure
$Spec_{adj}$	<b>0.795</b>	0.609	0.689
$Spec_{varg}$	0.663	0.702	0.682
$Spec_{coldoc}$	0.717	0.702	0.709
$Spec_{in}$	0.728	<b>0.907</b>	0.808
$Spec_{in/adj}$	<b>0.731</b>	<b>0.907</b>	<b>0.810</b>

## 5.2 Term Similarity

We evaluated similarity measures by comparing with taxonomy based similarity measure. (Budnitsky & Hirst, 2006) calculated correlation coefficients (CC) between human similarity ratings and the five WordNet based similarity measures. Among the five computational measures, (Leacock & Chodorow, 1998)’s method showed the highest correlation coefficients, even though all of the measures showed similar ranging from 0.74 to 0.85. This result means that taxonomy based similarity is highly correlated to human similarity ratings. We can indirectly evaluate our similarity measures by comparing to taxonomy based similarity measure, instead of direct comparison to human rating. If applied similarity measure is qualified, the calculated similarity will be highly correlated to taxonomy based similarity. Leacock and Chodorow proposed following formula for computing the scaled semantic similarity between terms  $t_1$  and  $t_2$  in taxonomy.

$$Sim_{LC}(t_1, t_2) = -\log \frac{len(t_1, t_2)}{2 \times \max_{t \in Taxonomy} depth(t)} \quad (12)$$

where the denominator includes the maximum depth of given taxonomy, and  $len(t_1, t_2)$  is number of edges in the shortest path between word  $t_1$  and  $t_2$  in the taxonomy.

Besides CC with ontology based similarity measures, recall of a similarity measures is also important evaluation factor. We defined recall of similarity measure,  $R_{sim}$ , as the fraction of the

term pairs that have similarity values by the given measuring method as Eq. (13).

$$R_{sim} = \frac{\# \text{ similarity measured term pairs}}{\# \text{ all possible term pairs}} \quad (13)$$

We also defined F-measure for a similarity measure,  $F_{sim}$ , as harmonic means of CC and  $R_{sim}$ . Because CC is a kind of precision,  $F_{sim}$  is overall measure of precision and recall.

We calculated term similarity between all possible term pairs in finance ontology using the measures described in section 3. Additionally we introduced new similarity measure  $Sim_{in/varg}$  which is combined similarity of  $Sim_{varg}$  and  $Sim_{in}$ .  $Sim_{varg}$  and  $Sim_{in}$  between two terms are harmonized to  $Sim_{in/varg}$ . We also calculated  $Sim_{LC}$  based on finance ontology, and calculated CC between  $Sim_{LC}$  and results of other measures. Figure 5 shows variation of CC and recall as threshold of similarity changes from 0.0 to 1.0 for five similarity measures. Threshold is directly proportional to CC and inversely proportional to recall in ideal case. We normalized all similarity values to [0.0, 1.0] in each measure. CC grows as threshold increases in  $Sim_{coldoc}$  and  $Sim_{varg}$  as we expected. CC of CSM measure,  $Sim_{csm}$ , increased as threshold increased and decreased when threshold is over 0.6. For example two terms ‘asset’ and ‘current asset’ are very similar to each other based on  $Sim_{LC}$  measure, because edge count between two terms is one in finance ontology. The former can be modified many adjectives such as ‘intangible’, ‘tangible’, ‘new’ and ‘estimated’, while the latter is rarely modified by other adjectives in corpus because it was already extended from ‘asset’ by adding adjective ‘current’. Therefore, semantically similar terms do not always have similar adjective distributions. CC between  $Sim_{in}$  and  $Sim_{LC}$  showed high curve in low threshold, but downed as threshold increased. Similarity value above 0.6 is insignificant, because it is hard to be over 0.6 using Eq. (10). For example, similarity between ‘executive board meeting’ and ‘board meeting’ is 0.8, the maximum similarity in our test set. The average of inside-word similarity is 0.41.

$Sim_{varg}$  showed higher recall than other measures. This means that verb-argument relation is more abundant than other features in corpus.  $Sim_{in}$  showed the lowest recall because we could get valid similarity using Eq. (10).  $Sim_{varg}$  showed higher F-measure when threshold is over 0.2. This result illustrate that verb-argument relation is adequate feature to similarity calculation.

The combined similarity measure,  $Sim_{in/varg}$ , complement shortcomings of  $Sim_{In}$  and  $Sim_{varg}$ .  $Sim_{In}$  showed high CC but low recall. Contrarily  $Sim_{varg}$  showed low CC but high recall.  $Sim_{in/varg}$  showed the highest F-measure.

### 5.3 Taxonomy learning

In order to evaluate our approach we need to assess how good the automatically learned taxonomies reflect a given domain. The goodness is evaluated by the similarity of automatically learned taxonomy to reference taxonomy. We used (Cimiano et al., 2005)'s ontology evaluation method in which lexical recall ( $LR_{Tax}$ ), precision ( $P_{Tax}$ ) and F-measure ( $F_{Tax}$ ) of learned taxonomy are defined based on the notion of taxonomy overlap.  $LR_{Tax}$  is defined as the ratio of number of common terms in learned taxonomy and reference taxonomy over number of terms in reference taxonomy.  $P_{Tax}$  is defined as ratio of taxonomy overlap of learned taxonomy to reference taxonomy.  $F_{Tax}$  is harmonic mean of  $LR_{Tax}$  and  $P_{Tax}$ .

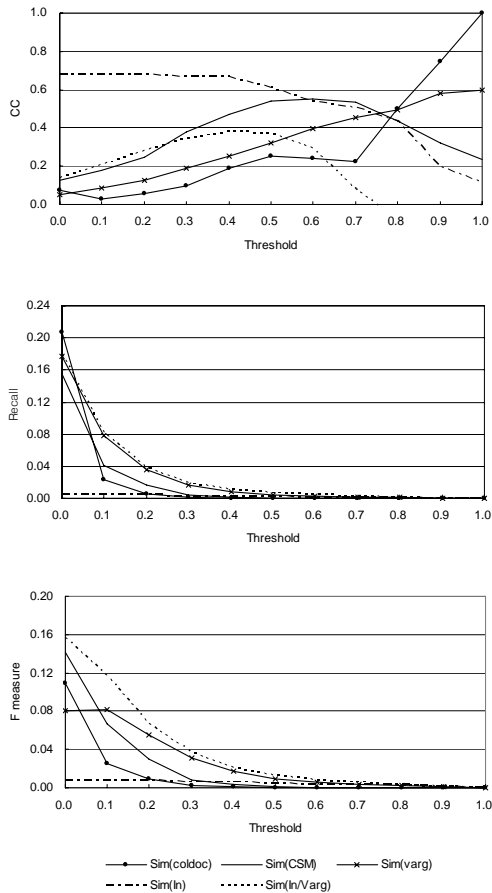


Figure 5 Correlation coefficient between  $Sim_{LC}$  and other similarity measures. Recall and F-measure of similarity measures

We generated four taxonomies,  $T_{coldoc}$ ,  $T_{csm}$ ,  $T_{fca}$ ,  $T_{spec/sim}$ , using four taxonomy learning methods: term co-occurring method, CSM method, FCA method and our method. We applied  $Spec_{in/adj}$  in specificity measuring and  $Sim_{in/varg}$  in similarity calculation because they showed the highest F-measure. In our method, the most probable one term was selected as hypernym of newly inserted term in each learning step.

Figure 6 shows variations of lexical recall, precision and F-measure of four methods as threshold changes. Threshold in each method represent different information to each other. Threshold in  $T_{csm}$  is variation of CSM values. Threshold in  $T_{coldoc}$  is variation of probability of two terms co-occur in a document. Threshold in  $T_{fca}$  is normalized frequency of contexts. Threshold in  $T_{spec/sim}$ , is variation of similarity.

$T_{spec/sim}$  showed the highest lexical recall. Lexical recall is tightly related to recall in similarity measures.  $Sim_{in/varg}$  showed the highest recall in similarity measures.  $T_{fca}$  and  $T_{csm}$  showed higher precision than other taxonomies. It is assumed that precision of taxonomy depends on

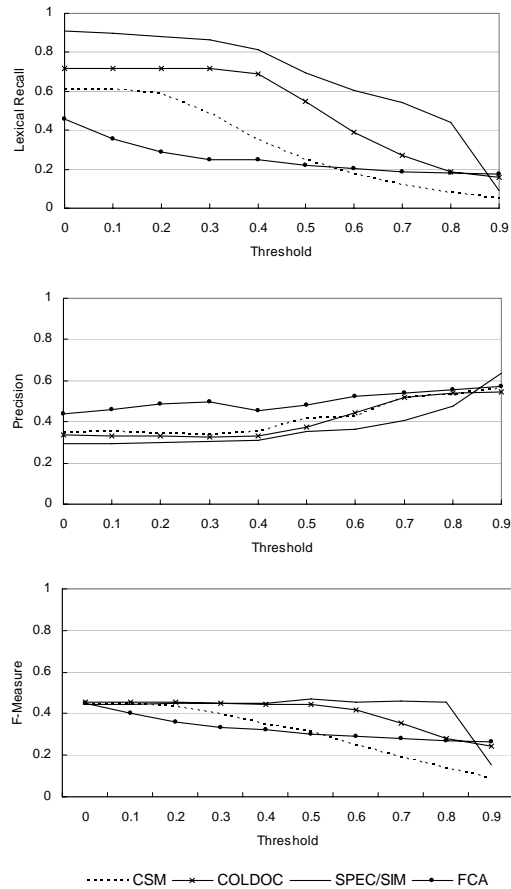


Figure 6. Lexical recall, precision and F-measure of taxonomy learning methods

the precision of specificity measures and the CC of similarity measures. In actual case,  $Sim_{\text{varg}}$  showed the most plausible curve in CC and  $Spe-c_{\text{adj}}$  showed the highest precision in specificity. Verb-argument relation and adjective-term relation are used in FCA and CSM methods respectively.  $T_{\text{spec/sim}}$  and  $T_{\text{coldoc}}$  showed higher F-measure curve than other two taxonomies due to high lexical recall. Although our method showed plausible F-measure, it showed the lowest precision. So other combination of similarity and specificity measures are needed to improve precision of learned taxonomy.

## 6 Conclusion

We have presented new taxonomy learning method with term similarity and specificity taken from domain-specific corpus. It can be applied to different domains as it is; and, if we have a syntactic parser available, to different languages. We analyzed the features used in previous researches in view of term specificity and similarity. In this analysis, we found that the features embed the characteristics of both conditions.

Compared to previous approaches, our method has advantages in that we can use different features for term specificity and similarity. It makes easy to analyze errors in taxonomy learning step, whether the wrong relations are caused by specificity errors or by similarity errors. The main drawback of our method, as it is now, is that the effect of wrong located terms in upper level propagates to lower levels.

Until now, researches on automatic ontology learning especially taxonomic relation showed very low precision. Human experts' intervention is inevitable in automatic learning process to make applicable taxonomy. Future work is to make new model where human experts and system work interactively in ontology learning process in order to balance cost and precision.

## Reference

S. Caraballo, E. Charniak. 1999. Determining the Specificity of Nouns from Text. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 63-70

P. Cimiano, A. Hotho, S.Staab. 2005. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of AI Research*, Vol. 24, pp. 305-339

M. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the*

*14th International Conference on Computational Linguistics*

- L. Iwanska, N. Mata and K. Kruger. 2000. Fully automatic acquisition of taxonomic knowledge from large corpora of texts. In Iwanska, L. & Shapiro, S. (Eds.), *Natural Language Processing and Knowledge Processing*, pp. 335-345, MIT/AAAI Press.
- E. Yamamoto, K. Kanzaki and H. Isahara. 2005. Extraction of Hierarchies Based on Inclusion of Co-occurring Words with Frequency Information. *Proceedings of 9th International Joint Conference on Artificial Intelligence*, pp. 1160-1167
- A. Burgun, O. Bodenreider. 2001. Aspects of the Taxonomic Relation in the Biomedical Domain, *Proceedings of International Conference on Formal Ontology in Information Systems*, pp. 222-233
- Mark Sanderson and Bruce Croft. 1999. Deriving concept hierarchies from text. *Proceedings of the 22th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 206-213, 1999
- Karen Sparck Jones. 1972. Exhaustivity and Specificity *Journal of Documentation* Vol. 28, Num. 1, pp. 11-21
- S.K.M. Wong, Y.Y. Yao. 1992. An Information-Theoretic Measure of Term Specificity, *Journal of the American Society for Information Science*, Vol. 43, Num. 1. pp.54-61
- ISO 704. 2000. *Terminology work-Principle and methods*. ISO 704 Second Edition
- A. Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Journal of Information Processing and Management*, vol. 39
- Alexander Budanitsky, Graeme Hirst. 2006 Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*. Vol. 32 NO. 1, pp. 13-47(35)
- Claudia Leacock, Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christian Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, pp. 265-283
- Pum-Mo Ryu, Key-Sun Choi. 2005. An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning, In P. Buitelaar et al. (eds.), *Ontology Learning from Text: Methods, Evaluation and Applications*, Vol. 123, Frontiers in Artificial Intelligence and Applications, IOS Press
- Farid Cerbah. 2000. Exogeneous and Endogeneous Approaches to Semantic Categorization of Unknown Technical Terms. *Proceedings of the 18th International Conference on Computational Linguistics*, vol. 1, pp. 145-151