Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 35-52

Automated Alignment and Extraction of a Bilingual Ontology for Cross-Language Domain-Specific Applications

Jui-Feng Yeh*, Chung-Hsien Wu*, Ming-Jun Chen* and Liang-Chih Yu*

Abstract

This paper presents a novel approach to ontology alignment and domain ontology extraction from two existing knowledge bases: WordNet and HowNet. These two knowledge bases are automatically aligned to construct a bilingual ontology based on the co-occurrence of words in a bilingual parallel corpus. The bilingual ontology achieves greater structural and semantic information coverage from these two complementary knowledge bases. For domain-specific applications, a domain-specific ontology is further extracted from the bilingual ontology using the island-driven algorithm and domain-specific corpus. Finally, domain-dependent terminology and axioms between domain terminology defined in a medical encyclopedia are integrated into the domain-specific ontology. In addition, a metric based on a similarity measure for ontology evaluation is also proposed. For evaluation purposes, experiments were conducted comparing an automatically constructed ontology with a benchmark ontology constructed by ontology engineers or experts. The experimental results show that the constructed bilingual domain-specific ontology mostly coincided with the benchmark ontology. As for application of this approach to the medical domain, the experimental results show that the proposed approach outperformed the synonym expansion approach to web search.

Keywords: Ontology, island driven algorithm, cross language application, WordNet, HowNet

* Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, ROC
E-mail: {jfyeh, chwu, mjchen, lcyu}@csie.ncku.edu.tw

1. INTRODUCTION

In the past few decades, a considerable number of studies been invested focused on developing concept bases for building technology that allows knowledge reuse and sharing. As information exchangeability and communication becomes increasingly global, multilingual lexical resources that provide transnational services are becoming increasingly important. On the other hand, multi-lingual ontologies are very important for natural language processing, such as machine translation (MT), web mining [Oyama *et al.* 2004], and cross-language information retrieval (CLIR). Generally, a multi-lingual ontology maps the keywords of one language to another language, or computes the co-occurrence of the words among languages. A key merit of a multilingual ontology is that it can achieve greater relation and structural information coverage by aligning or merging two or more language-dependent ontologies with different semantic features.

In recent years, significant effort has focused on constructing ontologies manually according to domain experts' knowledge. Manual ontology merging using conventional editing tools without intelligent support is difficult, labor intensive, and error prone. Therefore, several systems and frameworks to help knowledge engineers perform ontology merging have recently been proposed [Noy and Musen 2000]. To avoid reiteration in ontology construction, algorithms for ontology merging [UMLS http://umlsks.nlm.nih.gov] [Langkilde and Knight 1998] and ontology alignment [Vossen and Peters 1997] [Weigard and Hoppenbrouwers 1998] [Asanoma 2001] have been investigated. In these approaches, the final ontology is a merged version of the original ontologies with aligned links between them [Daudé *et al.* 2003]. Alignment is usually performed when ontologies cover domains that are complementary to each other. In the past, a domain ontology was usually constructed manually based on the knowledge or experience of experts or ontology engineers. Recently, automatic and semi-automatic methods have been developed. OntoExtract [Fensel *et al.* 2002] [Missikoff *et al*. 2002] provides an ontology engineering chain for constructing a domain ontology from WordNet and SemCor. Some recent approaches have been discussed in [Euzenat *et al.* 2004]. In [Euzenat *et al.* 2004], the alignment approaches were classified as local or global methods. Four main local methods, that is, the terminological, extensional, semantics, and structure methods, were introduced to measure the correspondence between two ontologies at the local level. Nowadays, much work is being invested in ontology construction for domain applications. Performing authoritative evaluation of ontologies is becoming a critical issue. Some evaluation methods are integrated into ontology tools to detect and prevent mistakes, which might be made in the course of developing taxonomies with frames as described in [Gómez-Pérez 2001]. They defined three main types of mistakes: inconsistency, incompleteness, and redundancy mistakes.

Although the previous research on ontology alignment has achieved much, some

important issues still require further investigation: (1) How can we to construct or extract domain concepts from a corpus? (2) Should the alignment of a cross-language or multilingual ontology be performed automatically or semi-automatically? (3) Authoritative assessment of ontology construction is desirable. In this study, the WordNet and HowNet knowledge bases were aligned to construct a bilingual universal ontology based on the co-occurrence of words in a bilingual parallel corpus. For domain-specific applications, the medical domain ontology was further extracted from the universal ontology using the island-driven algorithm and two corpora, one for the medical domain and another for the contrastive domain. Finally, axioms between medical terminology were derived based on a medical encyclopedia. A benchmark ontology based on the Unified Medical Language System (UMLS) and constructed by ontology engineers and experts was used to evaluate the constructed bilingual ontology. This paper also defines two measures, the taxonomic relation and non-taxonomic relation, as quantitative metrics for evaluating ontologies.

The rest of the paper is organized as follows. Section 2 describes the ontology construction process. Section 3 presents experimental results for the evaluation of our approach. Section 4 gives some concluding remarks.

2. Ontology Construction

Figure 1 shows a block diagram of the ontology construction process. There are two major stages in the proposed approach: bilingual ontology alignment and domain ontology extraction.

2.1 Bilingual Ontology Alignment

In this approach, a bilingual ontology is constructed by aligning Chinese words in HowNet with their corresponding synsets defined in WordNet according to the co-occurrence of the words in a bilingual parallel corpus. The hierarchical structure of the ontology is actually a conversion of HowNet. One of the important parts of HowNet consists of definitions of lexical entries. In HowNet, each lexical entry is defined as a combination of one or more primary features and a sequence of secondary features. The primary features indicate the entry's category, for example, the relation "is-a" in a hierarchical structure. Based on the entry's category, the secondary features make the entry's sense more explicit, but they are non-taxonomic. Totally, 1,521 primary features are divided into 6 upper categories: Event, Entity, Attribute Value, Quantity, and Quantity Value. These primary features are organized into a hierarchical structure.

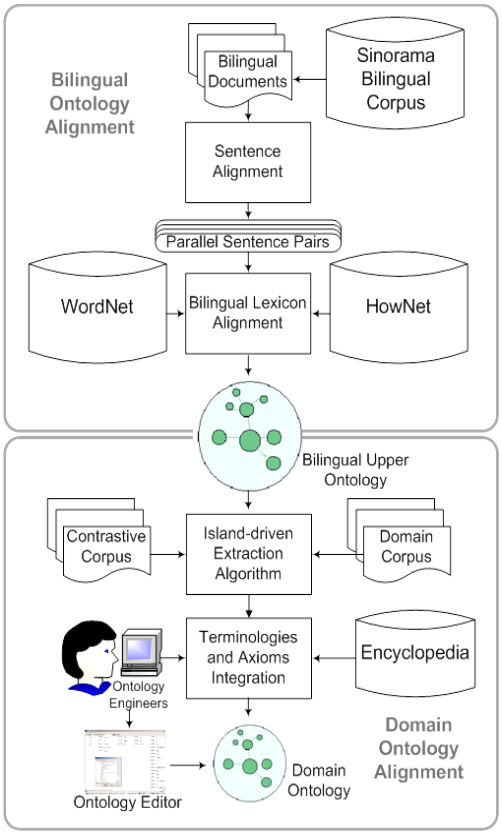


Figure 1. Ontology construction framework

In the alignment process, the Sinorama [Sinorama 2001] database, containing over 6,500 documents with 48,000,000 words from 1976 to 2000 in Chinese and English, is adopted as the bilingual parallel corpus. This corpus is then used to compute the conditional probability of the words in WordNet, given the words in HowNet. Then, a bottom up algorithm is used to perform relation mapping. In WordNet, a word may be associated with many synsets, each corresponding to a different sense of the word. To find a relation between two different words, all the synsets associated with each word are considered [Fellbaum 1998]. In HowNet, each word is composed of primary features and secondary features. The primary features indicate the word's category. The goal of this approach is to increase the amount of relation and structural information coverage by aligning their semantic features in WordNet and HowNet.

Equation (1) shows the alignment between the words in HowNet and the synsets in WordNet. Given a Chinese word, CW_i, the probability of the word being related to synset, $synset^k$, can be obtained via its corresponding English synonyms, EW_j^k, $j = 1, \ldots, m$, which

are the elements in $synset^k$. The probability is estimated as follows:

$$\begin{aligned}\Pr(synset^k \mid CW_i) &= \sum_{j=1}^{m} \Pr(synset^k, EW_j^k \mid CW_i) \\ &= \sum_{j=1}^{m} (\Pr(synset^k \mid EW_j^k, CW_i) \times (\Pr(EW_j^k \mid CW_i)),\end{aligned} \tag{1}$$

where

$$\Pr(synset^k \mid EW_j^k, CW_i) = \frac{N(synset_j^k, EW_j^k, CW_i)}{\sum_l N(synset_j^l, EW_j^k, CW_i)}. \tag{2}$$

In the above equation, $N(synset_j^k, EW_j^k, CW_i)$ represents the number of co-occurrences of CW_i, EW_j^k, and $synset_j^k$. The probability $\Pr(EW_j^k \mid CW_i)$ is set to one when at least one of the primary features, $PF_i^l(CW_i)$, of the Chinese word CW_i defined in HowNet matches one of the ancestor nodes of synset , $synset_j^k(EW_j)$, except for the root nodes in the hierarchical structures of the noun and verb. Otherwise, the probability $\Pr(EW_j^k \mid CW_i)$ is set to zero:

$$\Pr\left(EW_j \mid CW_i\right) = \begin{cases} 1, \ if \quad \left(\bigcup_l PF_i^l(CW_i) - \{entity, event, act, play\}\right) \cap \\ \qquad \left(\bigcup ancestor(\bigcup_k synset_j^k(EW_j)) - \{entity, event, act, play\}\right) \neq \varnothing, \\ 0, \quad Otherwise \end{cases} \tag{3}$$

where {entity, event, act, play} is the concept set in the root nodes of HowNet and WordNet, and $\left(\bigcup_l PF_i^l(CW_i) - \{entity, event, act, play\}\right)$ represents all the primary features of the Chinese word CW_i except for {entity, event, act, play}. Finally, the Chinese concept, CW_i, is integrated into the synset , $synset_j^k$, in WordNet as long as the probability, $\Pr(synset^k \mid CW_i)$, is not zero. Figure 2(a) shows the concept tree generated by aligning WordNet and HowNet.

2.2 Domain ontology extraction

Now, we will attempt to extend the ontology to domain applications. In domain-specific information retrieval, more detailed definitions and terminology are required. This paper proposes a two-stage domain ontology extraction method. This approach extracts the ontology from the cross-language ontology by using the island-driven algorithm in the first stage. The

terminology and axioms defined in a medical encyclopedia are integrated into the domain ontology in the second stage.

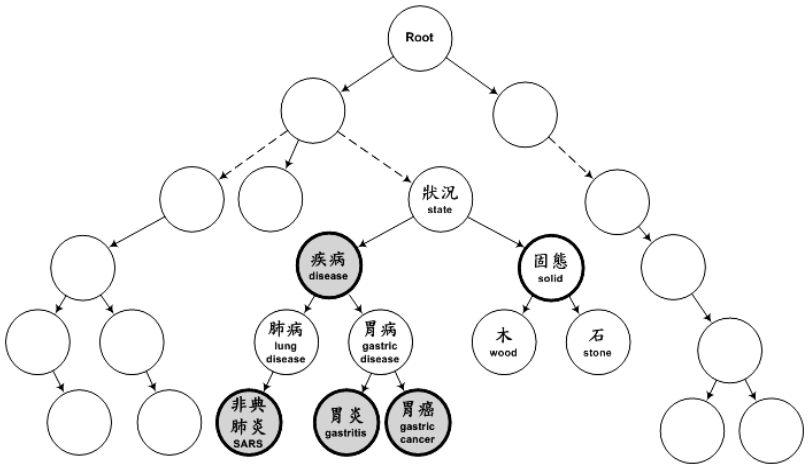


Figure 2(a). Concept tree generated by aligning WordNet and HowNet. The nodes in bold circles represent operative nodes following concept extraction. The nodes on gray backgrounds represent operative nodes following relationn expansion.

2.2.1 Extraction using the island-driven algorithm

Generally, an ontology provides consistent concepts and world representations necessary for clear communication within the knowledge domain. Even in domain-specific applications, the number of words can be expected to be huge. Synonym pruning is an effective way to perform word sense disambiguation. This paper proposes a corpus-based statistical approach to extracting a domain ontology. The steps are listed as follows:

Step 1. Linearization: In this step, the tree structure in the general purpose ontology shown in Figure 2(a) is decomposed into a vertex list that is an ordered node sequence starting at the root node and ending at the leaf nodes.

Step 2. Concept extraction from the corpus: The node is defined as an operative node when the tf-idf value of word W_i in the domain corpus is higher than that in its corresponding contrastive (out-of-domain) corpus. That is,

$$operative\_node(W_i) = \begin{cases} 1, & if \;\; tf\text{-}idf_{Domain}(W_i) > tf - idf_{Contrastive}(W_i) \\ 0, & Otherwiae \end{cases}, \tag{4}$$

where

$$tf-idf_{Domain}(W_i) = freq_{i,Domain} \times \log \frac{n_{i,Domain} + n_{i,Contrastive}}{n_{i,Domain}},$$

$$tf-idf_{Contrastive}(W_i) = freq_{i,Contrastive} \times \log \frac{n_{i,Domain} + n_{i,Contrastive}}{n_{i,Contrastive}}.$$

In the above equations, $freq_{i,Domain}$ and $freq_{i,Contrastive}$ are the frequencies of word W_i in the domain documents and its contrastive (out-of-domain) documents, respectively; $n_{i,Domain}$ and $n_{i,Contrastive}$ are the numbers of documents containing word W_i in the domain documents and its contrastive documents, respectively. The nodes shown in bold circles in Figure 2(a) represent operative nodes.

Step 3. Relation expansion using the island-driven algorithm: Some domain concepts are no longer operative after the previous steps have been performed due to the problem of data sparseness. According to the analysis performed during ontology construction, most of the inoperative concept nodes have operative hypernym nodes and hyponym nodes. Therefore, the island-driven algorithm is adopted to activate these inoperative concept nodes if their ancestors and descendants are all operative. The nodes shown on gray background in Figure 2(a) are activated operative nodes.

Step 4. Domain ontology extraction: In the final step, the linear vertex list sequence is merged into a hierarchical tree. However, some noisy concepts defined as nodes not belonging to this domain are operative according to Equation (5). For example, the node with the concept "solid" shown in Figure 2(b) is an operative noisy concept. Accordingly, the second goal is to filter out the nodes with operative noisy concepts. In this step, noisy concepts without ancestors or descendants belonging to the domain are removed. Finally, the domain ontology is extracted, and the final result is shown in Figure 2(b).

2.2.2 Axiom and terminology integration

In practice, specific domain terminology and axioms should be derived and introduced into an ontology for domain-specific applications. There are two approaches to integrating terminology and axioms into an ontology: the first one is manual editing performed by ontology engineers, and the second is automatic integration from a domain encyclopedia.

For medical domain applications, 1,213 axioms were derived here from a medical encyclopedia with terminology related to diseases, syndromes, and the clinic information. Figure 3 shows an example of an axiom. In this example, the disease "diabetes" is tagged as level "A," which means that this disease occurs frequently. The degrees for the corresponding syndromes indicate the causality between the disease and the syndromes. The axioms also provide two fields, "department of the clinical care" and "the category of the disease," for

medical information retrieval or other medical applications.

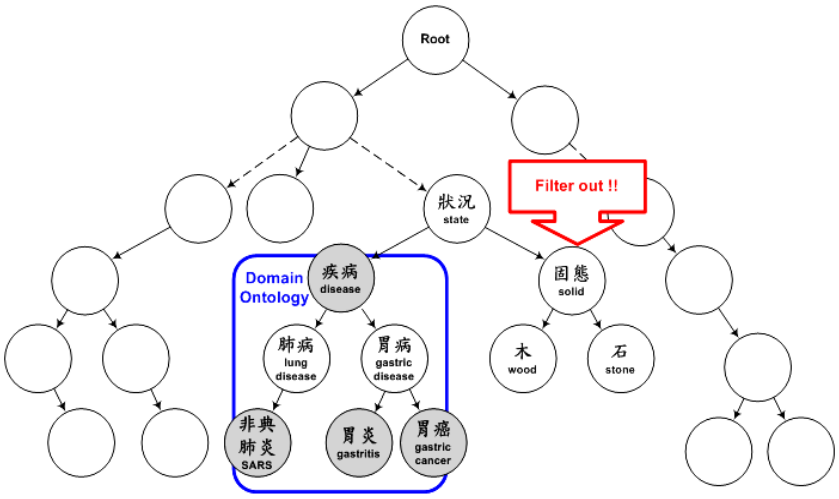


Figure 2(b). The domain ontology after isolated concepts are filtered out

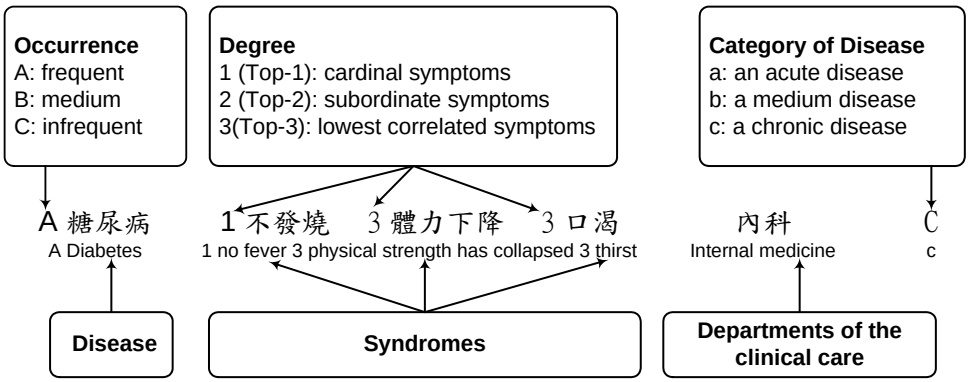


Figure 3. One example of an axiom

3. Evaluation

For quantitative evaluation of the ontology, two types of evaluation, conceptual evaluation and domain application evaluation, were adopted to evaluate the coincidence between the extracted domain ontology and the manually designed ontology. Furthermore, a medical web mining system was implemented to evaluate the practicability of the bilingual ontology.

3.1 Conceptual Evaluation

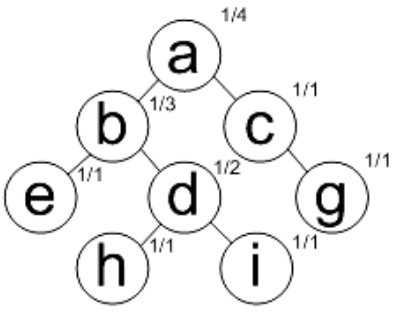
The benchmark ontology was created as a test-suite of reusable data which could be employed by ontology engineers for benchmarking purposes. The benchmark ontology was constructed by domain experts, including two doctors and one pharmacologist, based on the Unified Medical Language System (UMLS). The domain experts integrated the Chinese concepts without changing the contents of UMLS.

The construction of an ontology is generally evaluated using a two-layer measure, consisting of lexical and conceptual layers [Eichmann *et al.* 1998]. Evaluation in the conceptual layer seems to be more important than that in the lexical layer when the ontology is constructed by aligning or merging several well-defined source ontologies. There are two conceptual relation types of evaluation: taxonomic and non-taxonomic evaluation.

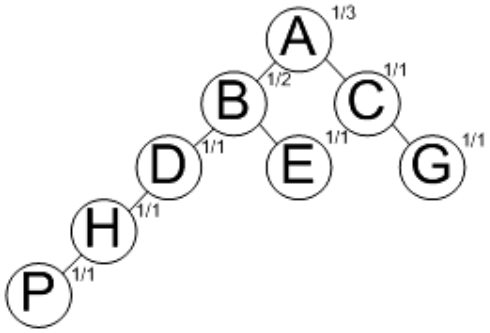
3.1.1 Evaluation of taxonomic relations

Evaluation of taxonomic relations is based not only on lexical similarity but also on hierarchical information according to the basic ontology definition. In this approach, obtaining the metric is a five-step process.

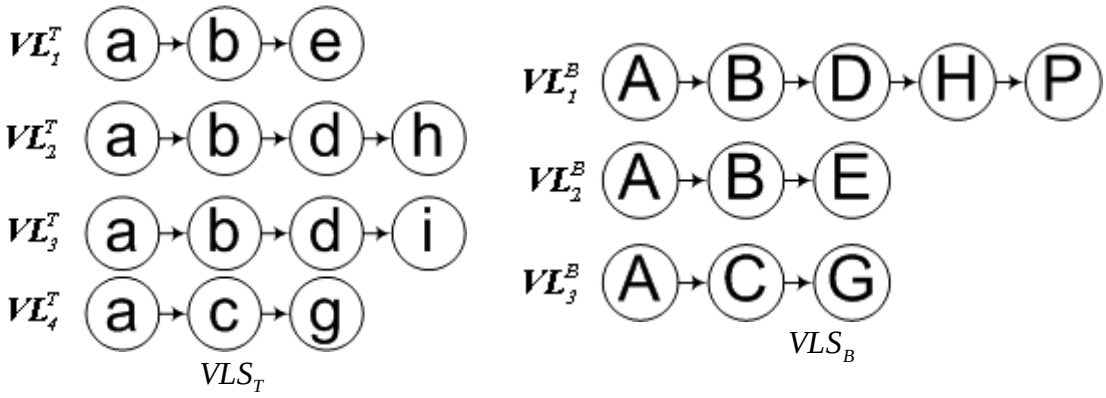
Step1. Linearization: In this step, the tree structure is decomposed into a vertex list as described in Section 2.2. The ontology, O_T, and the benchmark, O_B, are shown in Figures 4(a) and 4(b), respectively. After linearization is performed, the vertex list sets VLS_T and VLS_B are obtained as shown in Figure 4(c) and Figure 4(d), where $VLS_T = \{VL_1^T, ..., VL_p^T\}$; $VLS_B = \{VL_1^B, ..., VL_q^B\}$; VL_i^O represents the i-th vertex list of ontology *O*, and *p* and *q* are the numbers of vertex lists for the target ontology and the benchmark ontology, respectively.



(a)The taxonomic hierarchical representation of target ontology O_T



(b) The taxonomic hierarchical representation of benchmark ontology O_B



(c) The taxonomic vertex list representation of the target ontology

(d) The taxonomic vertex list representation of the benchmark ontology

Figure 4. Linearization of the target and benchmark ontologies

Step 2. Normalization: Since the frequencies of concepts in the vertex lists are not identical, normalization factors are introduced. For the target ontology, the set of factor vectors adopted for normalization is $NF^T = \{nf_1^T, nf_2^T, nf_3^T, nf_4^T, nf_5^T, ..., nf_m^T\}$, and for the benchmark ontology it is $NF^B = \{nf_1^B, nf_2^B, nf_3^B, nf_4^B, ..., nf_n^B\}$, where nf_i^o is the normalization factor for the i-th concept of ontology *O*. It is defined as the reciprocal of the number of vertex lists:

$$nf_i^O = \frac{1}{NV_i^O}, \tag{5}$$

where NV_i^O represents the number of vertex lists containing concept *i* in ontology *O*.

Step 3. Similarity estimation of two vertex lists: As the Figure 5 shows, the pairwise similarity of two vertex lists for the target ontology and benchmark ontology can be obtained using the Needleman/Wunsch techniques as described in the following steps:

1. Initialization: Create a matrix with *m*+1 columns and *n*+1 rows, where *m* and *n* are the numbers of nodes in the vertex lists of the target ontology and benchmark ontology, respectively. The first row and first column of the matrix can both be initially set to 0. That is,

$$Sim(m,n) = 0, \ if \ \ \mathrm{m} = 0 \ \ or \ \ \mathrm{n} = 0. \tag{6}$$

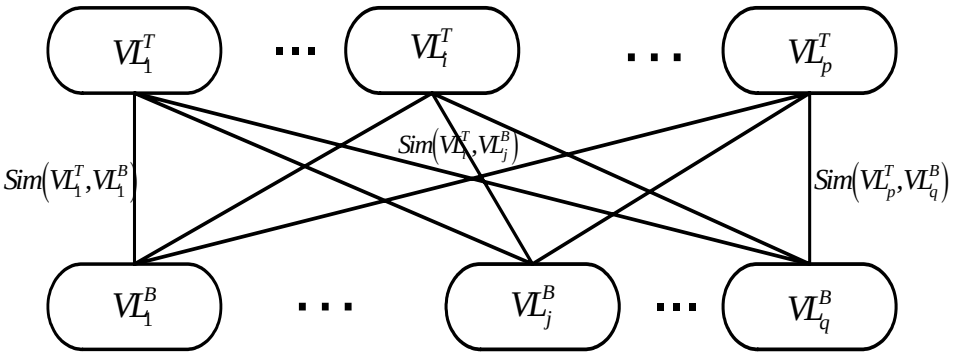


Figure 5. Pairwise similarity between the target ontology and benchmark ontology

2. Matrix filling: Assign values to the remaining elements in the matrix according to the following equation:

$$Sim(V_m^{T_i}, V_n^{B_j}) = \max \begin{cases} Sim(m-1, n-1) + \frac{1}{2}\left(nf_{m-1}^{T_i} + nf_{n-1}^{B_j}\right) \times Sim_{lexicon}\left(V_{m-1}^{T_i}, V_{n-1}^{B_j}\right), \\ Sim(m-1, n) + \frac{1}{2}\left(nf_{m-1}^{T_i} + nf_n^{B_j}\right) \times Sim_{lexicon}\left(V_{m-1}^{T_i}, V_n^{B_j}\right), \\ Sim(m, n-1) + \frac{1}{2}\left(nf_m^{T_i} + nf_{n-1}^{B_j}\right) \times Sim_{lexicon}\left(V_m^{T_i}, V_{n-1}^{B_j}\right). \end{cases} \quad (7)$$

There are some synonyms belonging to the same concept in one vertex. Thus, the lexical similarity can be defined as

$$Sim_{lexicon}\left(V_{m-1}^{T_i}, V_n^{B_j}\right) = \frac{\left|\text{Synonyms defined in } V_{m-1}^{T_i} \text{ and } V_n^{B_j}\right|}{\left|\text{Synonyms defined in } V_{m-1}^{T_i} \text{ or } V_n^{B_j}\right|}. \quad (8)$$

3. Traceback: Determine the actual alignment with the maximum score, $Sim(V_m^{T_i}, V_n^{B_j})$; therefore, the pairwise similarity is defined as follows:

$$Sim\left(VL_i^T, VL_j^B\right) \equiv \max_{m.n} Sim\left(V_m^{T_i}, V_n^{B_j}\right). \quad (9)$$

Step 4. Pairwise similarity matrix estimation: The pairwise similarity matrix is obtained after $p \times q$ iterations using the vertex list similarity defined in Step3. *p* and *q* are the numbers of vertex lists for the target ontology and benchmark ontology, respectively. Each element of the pairwise similarity matrix in Equation (10) is obtained from Equation (9):

$$PSM(O_T, O_B) \equiv \begin{bmatrix} Sim(VL_1^T, VL_1^B) & \cdots & Sim(VL_1^T, VL_q^B) \\ \vdots & \ddots & \vdots \\ Sim(VL_p^T, VL_1^B) & \cdots & Sim(VL_p^T, VL_q^B) \end{bmatrix}_{p \times q}. \quad (10)$$

Step 5. Evaluation of the taxonomic hierarchy: The total similarity between the target ontology and benchmark ontology, defined as the average similarity of all the vertex lists, is estimated as follows:

$$Sim_{taxonomic}(O_T, O_B) = \frac{1}{p} \sum_{i=1}^{p} \max_{1 \le j \le q} \left\{ Sim(VL_i^T, VL_j^B) \right\}. \quad (11)$$

3.1.2 Evaluation of non-taxonomic relations

Some relations defined in the ontology are non-taxonomic such as synonyms. In fact, lexical similarity is applied to measure the conceptual similarity. Lexical similarity is computed using the following equation:

$$Sim_{lexicon}\left(V_s^{T_i}, V_t^{B_j}\right) = \frac{\left|\text{Words defined in } V_s^{T_i} \text{ and } V_t^{B_j}\right|}{\left|\text{Words defined in } V_s^{T_i} \text{ or } V_t^{B_j}\right|}. \quad (12)$$

Therefore, evaluation of all of the whole non-taxonomic relations is performed according to the following equation:

$$Sim_{non-taxonomic}(O_T, O_B) = \frac{1}{p \times q} \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{s} \sum_{t} Sim_{lexicon}\left(V_s^{T_i}, V_t^{B_j}\right). \quad (13)$$

3.1.3 Evaluation results

Using the benchmark ontology and evaluation metrics described in the previous sections, we obtained the evaluation results shown in Table 1. The matching ratios between the constructed ontology and benchmark ontology were 57% and 68% for taxonomic and non-taxonomic relations, respectively. From the experimental results, the following phenomena were discovered: first, the number of words mapped to the same concept in the upper layer of the ontology was larger than that in the lower layer because the terminology usually appeared in the lower layer. Owing to the lack of an authoritative benchmark, the metrics could not provide an ideal measure. The main weakness was the difference between the target and benchmark ontologies, especially the terminology used. Introducing concept or word frequency measures may lead to a significant improvement.

Table 1. Matching ratio between the target ontology and benchmark ontology

Taxonomic relation matching ratio	***57%***
Non-Taxonomic relation matching ratio	***68%***

3.2 Evaluation of domain application

To assess the performance of the ontology, a cross-language medical domain web-mining system was implemented. For domain concept extraction, a corpus was collected from several websites. A total of 2,322 web pages were collected as a medical domain corpus, and 8,133 web pages as a contrastive domain corpus. Besides the training corpus, 1,212 web pages different from the training sets and the test queries were also collected for the purpose of system evaluation. Forty users, who did not take part in system development, were asked to provide a set of queries given the collected web pages. After post-processing was performed, the duplicate queries and the queries that were out of the medical domain were removed. Finally, 3,207 test queries using natural language were obtained.

The baseline system is based on the Vector-Space Model (VSM). That is, a sequence of words is treated as a bag of words regardless of the word order. For a word sequence from a user's input, $q = \{q_1, q_2, \cdots\cdots q_n\}$, and a word sequence in a web page, $d = \{d_1, d_2, \cdots d_n\}$, the similarity is defined as the cosine function as follows:

$$Sim_{VSM}(D_i, q) = \cos(d, q) = \frac{\sum_{i=1}^{n} d_i q_i}{\sqrt{\sum_{i=1}^{n} d_i^2} \times \sqrt{\sum_{i=1}^{n} q_i^2}}, \tag{14}$$

where D_i is the i-th document in the web page and q is the user's query. This approach to key term expansion based on a synonym set is also adopted in the baseline system.

The conceptual relations and axioms defined in the medical ontology were integrated into the baseline as the ontology-based system. The medical web search engine was developed based on the constructed medical domain ontology consists of a relation inference module and axiom inference module. The functions of and techniques used with these modules are described in the following.

3.2.1 Relation inference module

For semantic representation, traditionally, keyword-based systems face two problems. First, ambiguity usually results from the polysemy of words. The domain ontology gives clear descriptions of the concepts. In addition, not all of the synonyms of a word should be expanded without any constraints being applied. Secondly, the relations between the concepts should be expanded and weighted in order to include more semantic information for semantic inferences. We treat each user's input and the content of a web page as a sequence of words.

The similarity between an input query and a web page is defined as the similarity between the two bags of words based on key concepts in the ontology [Yeh *et al.* 2004].

$$Sim_{relation}(D_i, q) = Sim_{relation}(d, q) = Sim_{relation}(d_1, d_2, \ldots, d_L, q_1, q_2, \ldots q_k)$$
$$= \begin{cases} 1 & d_l \text{ and } q_k \text{ are identical} \\ \sum_{k=1}^{K}\sum_{l=1}^{L}\left(\frac{1}{2^r}\right) & d_l \text{ and } q_k \text{ are hypernyms and } r \text{ is the number of levels in between} \\ \sum_{k=1}^{K}\sum_{l=1}^{L}\left(1-\frac{1}{2^t}\right)^2 & d_l \text{ and } q_k \text{ are synonyms and } t \text{ is the number of their common concepts} \\ 0 & \text{Other} \end{cases} \quad (15)$$

3.2.2 Axiom inference module

Some axioms, such as "result in" and "result from," that are expected to affect the performance of a web search system in a medical domain are defined in order to describe the relationships between syndromes and diseases. We collected data about syndromes and diseases from a medical encyclopedia and tagged the diseases with three levels according to their frequency of occurrence and tagged syndromes with four levels according to their significance with respect to a specific disease. The "result in" relation score is defined as $RI(D_i, q)$ if a disease occurs in the input query and its corresponding syndromes appear in the web page. Similarly, if a syndrome occurs in the input query and its corresponding disease appears in the web page, the "result from" relation score is defined as $RF(D_i, q)$. The relation score is estimated as described in [Yeh *et al.* 2004]:

$$\begin{aligned} Axiom(D_i, q) &= \max\{RI(D_i, q), RF(D_i, q)\} \\ &= \max\{RI(d_1, d_2, \ldots, d_p, q_1, q_2, \ldots, q_R), RF(d_1, d_2, \ldots, d_p, q_1, q_2, \ldots q_R)\} \\ &= \max\left\{\sum_{p=1,r=1}^{P,R} a_{pr}^{RI}, \sum_{p=1,r=1}^{P,R} a_{pr}^{RF}\right\}, \end{aligned} \quad (16)$$

where $a_{pr}^{RI} = 1/2^{n-1}$ if disease d_p results in syndrome q_r and q_r is the top-n feature of d_p. Similarly, $a_{pr}^{RF} = 1/2^{n-1}$ if syndrome d_p results from disease q_r and d_p is the top-n feature of q_r. The similarity between the i-th web page and query *q* is defined as

$$Sim_{axiom}(D_i, q) = \frac{Axiom(D_i, q)}{\sum_i Axiom(D_i, q)}. \quad (17)$$

3.2.3 Weight determination using the 11-avgP score

The medical domain web search system is modelled using a linear combination of a relational inference model and axiom inference model. The normalized weight factor, α, is employed for the purpose of concept expansion as follows:

$$Sim(D_i, q) = (1-\alpha)Sim_{relation}(D_i, q) + \alpha \times Sim_{axiom}(D_i, q). \quad (18)$$

An experiment was conducted to evaluate the estimation of the combination weights for each model. The results are shown in Figure 6. A performance measure called 11-AvgP [Eichmann and Srinivasan 1998] was used to summarize the precision and recall rates. The best 11-AvgP score was obtained when the weight α was set to 0.428.

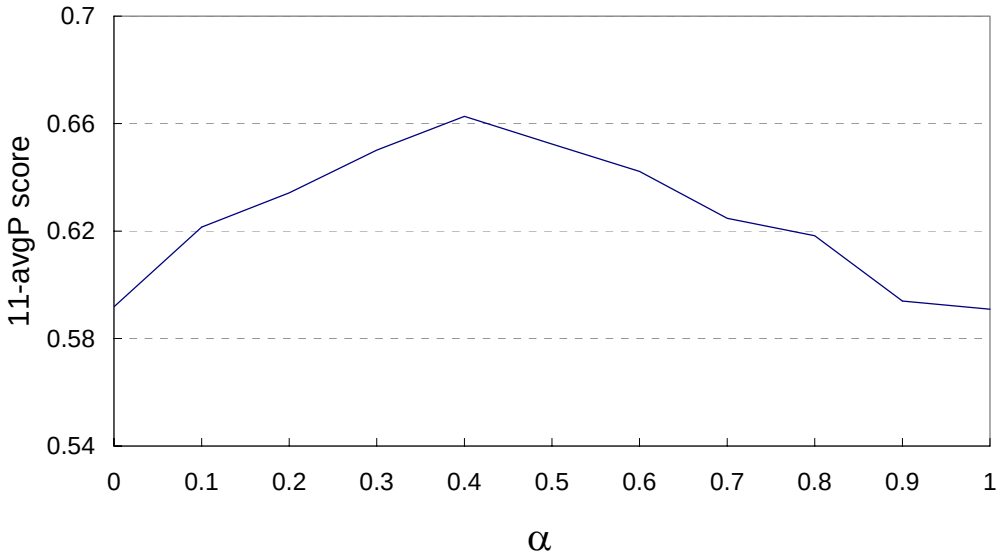


Figure 6. The 11-avgP score with different values of α

3.2.4 Evaluation of different inference modules

In the following experiments, web pages were separately evaluated by focusing on one inference module based on the domain-specific ontology at a time. That is, the mixture weight was set to 1 for one inference module, and the other weight was set to 0 in each evaluation. For comparison purposes, the keyword-based VSM approach and the ontology-based system were also evaluated, and the results are shown in Figure 7. The precision and recall rates were used as the evaluation measures. The ontology-based approach combines of concept inferences and axiom inferences as described in the previous sections. The results shown in Table 2 reveal that the ontology-based system outperformed the baseline system in synonym

expansion. Instead of keywords, the concepts defined in the ontology play an important role in term expansion for a specific domain. In addition, relation axioms are important and can be effectively used in domain applications; that is to say, the inference axioms provide semantic relationships between words.

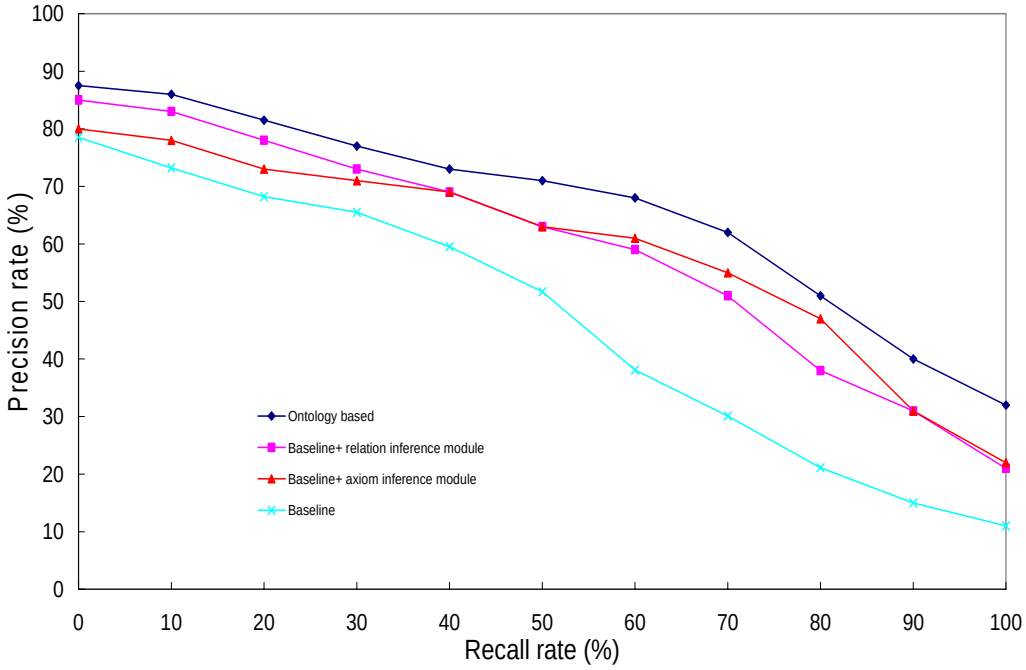


Figure 7. The precision rates and recall rates achieved with the proposed method and the baseline system

Table 2. Precision rates (%) at the 11-point recall level

Recall Level	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
Baseline system (Precision)	*78*	*73*	*68*	*65*	*60*	*52*	*38*	*30*	*21*	*15*	*11*
Ontology based (Precision)	*87*	*86*	*82*	*77*	*73*	*71*	*68*	*62*	*51*	*40*	*32*

4. CONCLUSIONS

A novel approach to automated ontology alignment and domain ontology extraction from two knowledge bases has been presented in this paper. In this study, a bilingual ontology has been developed from two well established knowledge bases, WordNet and HowNet, based on the co-occurrence of words in a parallel bilingual corpus. A domain-dependent ontology has been further extracted from the universal ontology using the island-driven algorithm and a domain corpus as well as a contrastive corpus. In addition, domain-specific terms and axioms have also been added to the domain ontology. A metric based on the similarity measure for ontology evaluation has also been proposed. The experimental results show that the proposed approach can extract an aligned bilingual domain-specific ontology which mostly coincides with a corresponding manually designed ontology. We have also applied the obtained domain-specific ontology to web page search in a medical domain. The experimental results show that the proposed approach outperformed the synonym expansion approach.

REFERENCES

Asanoma, N., "Alignment of Ontologies: WordNet and Goi-Taikei," *Proc. of WordNet and Other Lexical Resources Workshop Program,* NAACL2001, 2001, pp. 89-94

Daudé, J., L. Padró. and G. Rigau, "Validation and Tuning of Wordnet Mapping Techniques," *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP'03).* Borovets, Bulgaria, 2003.

Eichmann, D., M. Ruiz and P. Srinivasan, "Cross-language information retrieval with the UMLS Metathesaurus," *Proc. of ACM Special Interest Group on Information Retreival (SIGIR),* ACM Press, NY (1998), 1998, pp. 72-80.

Euzenat, J., T. Le Bach, J. Barrasa, P. Bouquet, J. De Bo, R. Dieng, M. Ehrig, M. Hauswirth, M. Jarrar, R. Lara, D. Maynard, A. Napoli, G. Stamou, H. Stuckenschmidt, P. Shvaiko, S. Tessaris, S. Van Acker and I. Zaihrayeu, State of the Art on Ontology Alignment. Knowledge Web Deliverable D2.2.3, INRIA, Saint Ismier, 2004.

Fensel, D., C. Bussler, Y. Ding, V. Kartseva, M. Klein, M. Korotkiy, B. Omelayenko and R. Siebes, "Semantic Web Application Areas," *Proc. of the 7th International Workshop on Applications of Natural Language to Information Systems*, Stockholm - Sweden, June 27--28, 2002.

Fellbaum, F. C., "WordNet an electronic Lexical Database," The MIT Press, 1998. pp. 307-308

Gómez-Pérez, A., "Evaluating ontologies: Cases of Study," *IEEE Intelligent Systems and their Applications: Special Issue on Verification and Validation of ontologies*. vol. 16, no 3. March 2001, pp. 391-409.

HowNet, http://www.keenage.com/

Langkilde, I. and K. Knight, "Generation that Exploits Corpus-Based Statistical Knowledge," *Proc. of COLING-ACL* 1998, pp. 704-710

Levensthein, V., "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics–Doklady*, vol.10, no. 8, 1966, pp.707–710.

Missikoff, M., R. Navigli and P. Velardi, "An Integrated Approach for Web Ontology Learning and Engineering," *IEEE Computer*, November 2002, pp. 60-63.

Noy, N. F. and M. Musen, "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment," *Proc. of the National Conference on Artificial Intelligence*. AAAI2000. 2000, pp.450-455

Oyama, S., T. Kokubo and T. Ishida, "Domain-Specific Web Search with Keyword Spice," *IEEE Transactions on Knowledge and Data Engineering*,vol 16, nO. 1, 2004, pp. 17-27.

Sinorama Magazine and Wordpedia.com Co.. Multimedia CD-ROMs of Sinorama from 1976 to 2000, Taipei, 2001.

UMLS, http://www.nlm.nih.gov/research/umls/

Vossen, P. and W. Peters, "Multilingual design of EuroWordNet," *Proc. of the Delos workshop on Cross-language Information Retrieval.* 1997.

Weigard, H. and S. Hoppenbrouwers, "Experiences with a multilingual ontology-based lexicon for news filtering," *Proc. of the 9th International Workshop on Database and Expert Systems Applications*. 1998, pp.160-165.

WordNet, http://www.cogsci.princeton.edu/~wn/

Yeh, J.F., C.H. Wu, M.J. Chen and L.C. Yu, "Automated Alignment and Extraction of Bilingual Domain Ontology for Medical Domain Web Search," *Proc. Of the 3rd SIGHAN Workshop on Chinese Language Learning*, ACL2004, Barcelona, 2004, pp.65-71