Automated Alignment and Extraction of Bilingual Domain Ontology for Medical Domain Web Search

Jui-Feng Yeh[*+] , Chung-Hsien Wu[*], Ming-Jun Chen[*] and Liang-Chih Yu[*]
[*]Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan, R.O.C.
{chwu, jfyeh, mjchen,lcyu}@csie.ncku.edu.tw
[+] Department of Computer Application Engineering
Far East College, Taiwan, R.O.C.

Abstract

This paper proposes an approach to automated ontology alignment and domain ontology extraction from two knowledge bases. First, WordNet and HowNet knowledge bases are aligned to construct a bilingual universal ontology based on the co-occurrence of the words in a parallel corpus. The bilingual universal ontology has the merit that it contains more structural and semantic information coverage from two complementary knowledge bases, WordNet and HowNet. For domain-specific applications, a medical domain ontology is further extracted from the universal ontology using the island-driven algorithm and a medical domain corpus. Finally, the domain-dependent terms and some axioms between medical terms based on a medical encyclopaedia are added into the ontology. For ontology evaluation, experiments on web search were conducted using the constructed ontology. The experimental results show that the proposed approach can automatically align and extract the domain-specific ontology. In addition, the extracted ontology also shows its promising ability for medical web search.

1 Introduction

In intelligent mining, in order to obviate the unnecessary keyword expansion, some knowledge base should be involved in the intelligent information system. In recent years, considerable progress has been invested in developing the conceptual bases for building technology that allows knowledge reuse and sharing. As information exchangeability and communication becomes increasingly global, multiple-language lexical resources that can provide transnational services are becoming increasingly important. Over the last few years, significant effort has been made to construct the ontology manually according to the domain expert's knowledge. Manual ontology merging using conventional editing tools without intelligent support is difficult, labor intensive and error prone. Therefore, several systems and frameworks for supporting the knowledge engineer in the ontology merging task have recently been proposed. To avoid the reiteration in ontology construction, the algorithm of ontology merge (UMLS http://umlsks.nlm.nih.gov/) (Langkilde and Knight 1998) and ontology alignment (Vossen and Peters 1997) (Weigard and Hoppenbrouwers 1998) (Asanoma 2001) were invested. The final ontology is a merged version of the original ontologies. The two original ontologies persist, with links established between them in alignment. Alignment usually is performed when the ontologies cover domains that are complementary to each other. In the past, domain ontology was usually constructed manually according to the knowledge or experience of the experts or ontology engineers. Recently, automatic and semi-automatic methods have been developed. OntoExtract (Fensel et al. 2002) (Missikoff et al. 2002) provided an ontology engineering chain to construct the domain ontology from WordNet and SemCor.

On the other hand, multi-lingual ontology is very important for natural language processing, such as machine translation (MT), web mining and cross language information retrieval (CLIR). Generally, a multi-lingual ontology maps the keyword set of one language to another language, or compute the co-occurrence of the words among languages. In addition, a key merit for multilingual ontology is that it can increase the relation and structural information coverage by aligning two or more language-dependent ontologies with different semantic features.

Nowadays large collections of information in various styles are available on the Internet. And finding desired information on the World Wide Web is becoming a critical issue. Some general-purpose search engine like Google (http://www.google.com) and Altavista (http://www.altavista.com/) provide the facility to

mine the web. There are three major research areas about web mining: web content mining, web structure mining and web usage mining. This paper proposes a novel method to web content mining with unstructured web pages. There are many approaches in the view of natural language processing. According to the representation of web pages, there are three kinds of the content: bag of words (with order or not) (Kargupta et al. 1997) (Nahm and Mooney, 2000), phrases (Ahonen et al. 1998) (Frank et al. 1999)(Yang et al. 1999), relational terms (Cohen, 1998) (Junker 1999) and concept categories. We proposed an ontology-based web search approach. Unfortunately, there are some irrelevant pages obtained and these pages result in low precision rate and recall rate due to the problem of polysemy. To solve this problem, domain knowledge becomes necessary. The domain-specific web miners like SPIRAL, Cora (Cohen, 1998), WebKB (Martin and Eklund 2000) and HelpfulMed (Chen et al. 2003) are employed as the special search engine for the interesting topic. These ones dedicated to recipes are less likely to return irrelevant web pages when the query is entered.

In this paper, WordNet and HowNet knowledge bases are aligned to construct a bilingual universal ontology based on the co-occurrence of the words in a parallel corpus. For domain-specific applications, a medical domain ontology is further extracted from the universal ontology using the island-driven algorithm (Lee et al. 1995) and a medical domain corpus. Finally, the axioms between medical terms are derived based on semantic relations. A web search system for medical domain based on the extracted domain ontology is realized to demonstrate the feasibility of the methods proposed in this paper.

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

2 Methodologies

Figure 1 shows the block diagram for ontology construction and the framework of the domain-specific web search system. There are four major processes in the proposed system: bilingual ontology alignment, domain ontology extraction, knowledge representation and domain-specific web search.

2.1 Bilingual Ontology Alignment

In this approach, the cross-lingual ontology is constructed by aligning the words in WordNet with their corresponding words in HowNet. First, the Sinorama (Sinorama 2001) database is adopted as the bilingual language parallel corpus to compute the conditional probability of the words in WordNet, given the words in HowNet. Second, a bottom up algorithm is used for relation mapping.

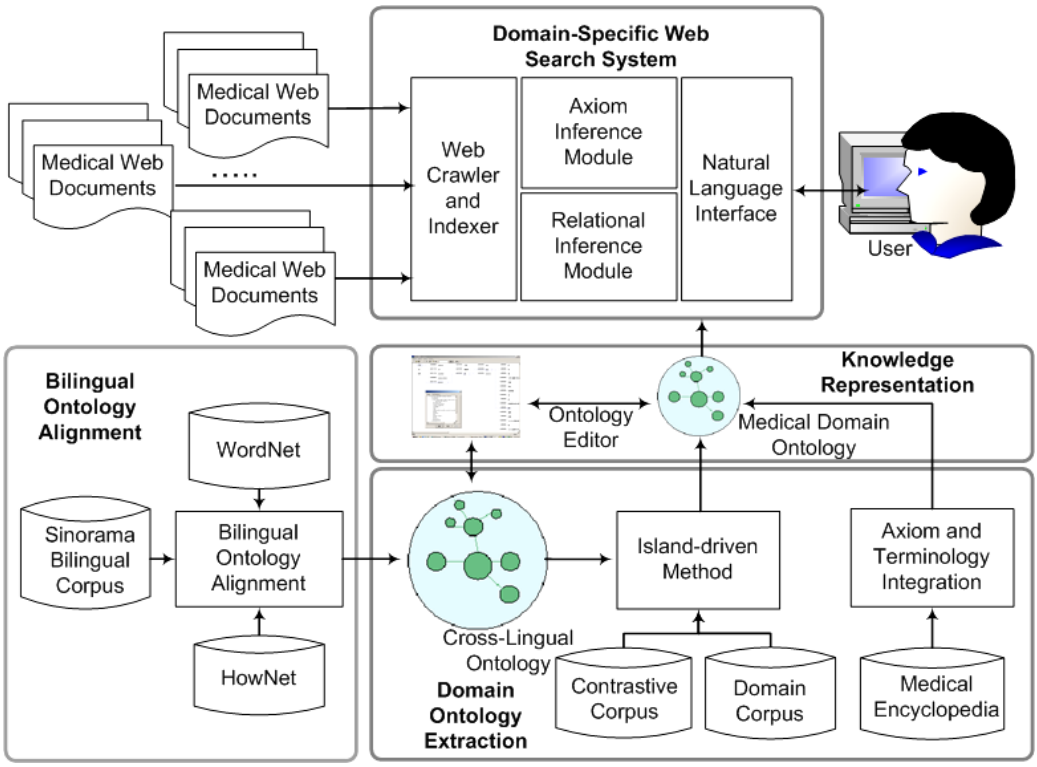

Figure 1 Ontology construction framework and the domain-specific web search system

In WordNet a word may be associated with many synsets, each corresponding to a different sense of the word. When we look for a relation between two different words we consider all the synsets associated with each word (Christiane 1998). In HowNet, each word is composed of primary features and secondary features. The primary features indicate the word's category. The purpose of this approach is to increase the relation and structural information coverage by aligning the above two language-dependent ontologies, WordNet and HowNet, with different semantic features.

The relation "is-a" defined in WordNet corresponds to the primary feature defined in HowNet. Equation (1) shows the mapping between the words in HowNet and the synsets in WordNet. Given a Chinese word, CW_i , the probability of the word related to synset, $synset^k$, can be obtained via its corresponding English synonyms, EW_j^k, $j=1,...,m$, which are the elements in $synset^k$. The probability is estimated as follows.

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

where

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

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

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

where {enitity,event,act,play} is the concept set in the root nodes of HowNet and WordNet.

Finally, the Chinese concept, CW_i , has been 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 shows the concept tree generated by aligning WordNet and HowNet.

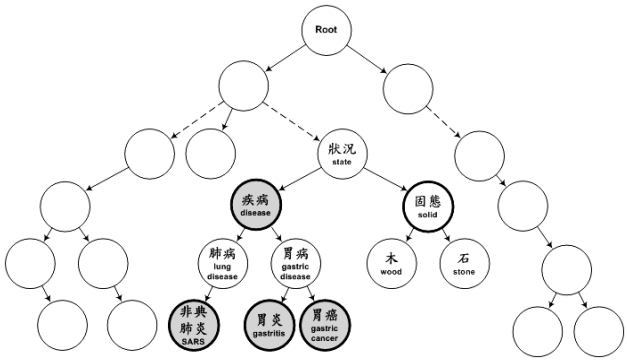

Figure 2. Concept tree generated by aligning WordNet and HowNet. The nodes with bold circle represent the operative nodes after concept extraction. The nodes with gray background represent the operative nodes after relation expansion.

2.2 Domain ontology extraction

There are two phases to construct the domain ontology: 1) extract the ontology from the cross-language ontology by island-driven algorithm, and 2) integrate the terms and axioms defined in a medical encyclopaedia into the domain ontology.

2.2.1 Extraction by island-driven algorithm

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 numerous. Synonym pruning is an effective alternative to word sense disambiguation. This paper proposes a corpus-based statistical approach to extracting the domain ontology. The steps are listed as follows:

Step 1: Linearization: This step decomposed the tree structure in the universal ontology shown in Figure 2 into the vertex list that is an ordered node sequence starting at the leaf nodes and ending at the root node.

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,

$$operative\_node(W_i) = \begin{cases} 1, & if\ \ Tf-idf_{Domain}(W_i) > Tf-idf_{Contrastive}(W_i) \\ 0, & Otherwise \end{cases} \quad (4)$$

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 the documents containing word W_i in the domain documents and its contrastive documents, respectively. The nodes with bold circle in Figure 2 represent the operative nodes.

Step 3: Relational expansion using the island-driven algorithm: There are some domain concepts not operative after the previous steps due to the problem of insufficient data. From the observation in 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 with gray background shown in Figure 2 are the activated operative nodes.

Step 4: Domain ontology extraction: The final step is to merge the linear vertex list sequence into a hierarchical tree. However, some noisy concepts not belonging to this domain ontology are operative after step 3. These noisy nodes with inoperative noisy concepts should be filtered out automatically. Finally, the domain ontology is extracted and the final result is shown in Figure 3.

After the above steps, a dummy node is added as the root node of the domain concept tree.

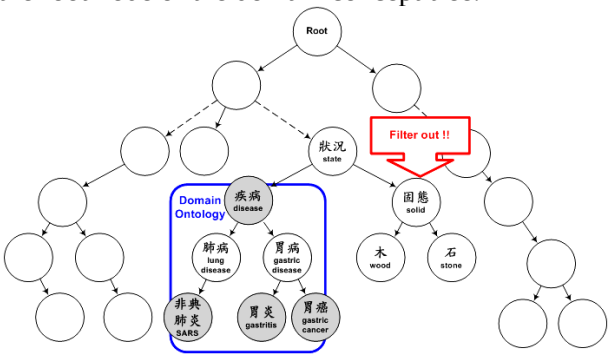

Figure 3 The domain ontology after filtering out the isolated concepts

2.2.2 Axiom and terminology integration

In practice, specific domain terminologies and axioms should be derived and introduced into the ontology for domain-specific applications. In our approach, 1213 axioms derived from a medical encyclopaedia have been integrated into the domain ontology. Figure 4 shows an example of the axiom. In this example, the disease "diabetes" is tagged as level "A" which represents that this disease is frequent in occurrence. And the degrees for the corresponding syndromes represent 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.

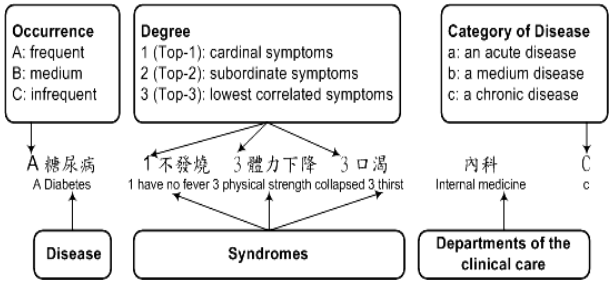

Figure 4 axiom example

2.3 Domain-specific web search

This paper proposed a medical web search engine based on the constructed medical domain ontology. The engine consists of natural language interface, web crawler and indexer, relation inference module and axiom inference module. The functions and techniques of these modules are described as follows.

2.3.1 Natural language interface and web crawler and indexer

Natural language interface is generally considered as an enticing prospect because it offers many advantages: it would be easy to learn and easy to remember, because its structure and vocabulary are already familiar to the user; it is particularly powerful because of the multitude of ways in which to accomplish a search action by using the natural language input. A natural language query is transformed to obtain the desired representation after the word segmentation, removing the stop words, stemming and tagging process.

The web crawler and indexer are designed to seek medical web pages from Internet, extract the content and establish the indices automatically.

2.3.2 Concept inference module

For semantic representation, traditionally, the keyword-based systems will introduce two problems. First, ambiguity usually results from the polysemy of words. The domain ontology will give a clear description of the concepts. In addition, not all the synonyms of the word should be expanded without constraints. Secondly, relations between the concepts should be expanded and weighted in order to include more semantic information for semantic inference. We treat each of the user's input and the content of web pages as a sequence of words. This means that the sequence of words is treated as a bag of words regardless of the word order. For the word sequence of the user's input,

$$q = W_q = w_{q1}, w_{q2}, \ldots, w_{qK},$$

and the word sequence of the web page,

$$A = W_A = w_{A1}, w_{A2}, \ldots, w_{AL},$$

The similarity between input query and the page is defined as the similarity between the two bags of words. The similarity measure based on key concepts in the ontology is defined as follows.

$$\begin{aligned} & Sim_{relation}(A_i, q) \\ &= Sim_{relation}(W_{A_i}, W_q) \\ &= Sim_{relation}(w_{A_i 1}, w_{A_i 2}, \ldots, w_{A_i L}, w_{q1}, w_{q2}, \ldots, w_{qK}) \\ &= \sum_{k=1,l=1}^{K,L} H_{kl} \end{aligned} \tag{5}$$

where H_{kl} is concept similarity of w_{Al} and w_{qk}. Most of the keyword expansion approaches use the extension of scope by the synonyms. In this paper the similarity, H_{kl}, is defined as

$$H_{kl} = \begin{cases} 1 & w_{Al} \text{ and } w_{qk} \text{ are identical} \\ \dfrac{1}{2^r} & w_{Al} \text{ and } w_{qk} \text{ are hypernyms,} \\ & r \text{ is the number of levels in between} \\ \left(1 - \dfrac{1}{2^r}\right)^2 & w_{Al} \text{ and } w_{qk} \text{ are synonyms,} \\ & r \text{ is the number of their common concepts} \\ 0 & \text{others} \end{cases} \tag{6}$$

2.3.3 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 technical domain are defined to describe the relationship between syndromes and diseases. This aspect is the use of specific terms used in the medical domain. We collected the data about syndromes and diseases from a medical encyclopedia and tagged the diseases with three levels according to its occurrence and syndromes with four levels according to its significance to the specific disease. The "result in" relation score is defined as $RI(A_i, q)$ if a disease occurs in the input query and its corresponding syndromes appear in the web page. Similarly, if syndrome occurs in the input query and its corresponding disease appears in the web page, the "result from" relation score is defined as $RF(A_i, q)$. The relation score is estimated as follows.

$$\begin{aligned} Axiom(A_i, q) &= \max\{RI(A_i, q), RF(A_i, q)\} \\ &= \max\{RI(w_{A_i 1}, w_{A_i 2}, \ldots, w_{A_i P}, w_{q1}, w_{q2}, \ldots, w_{qR}), \\ &\qquad RF(w_{A_i 1}, w_{A_i 2}, \ldots, w_{A_i P}, w_{q1}, w_{q2}, \ldots, w_{qR})\} \\ &= \max\{\sum_{p=1,r=1}^{P,R} d_{pr}^{RI}, \sum_{p=1,r=1}^{P,R} d_{pr}^{RF}\}, \end{aligned} \tag{7}$$

where $d_{pr}^{RI} = 1/2^{n-1}$ if disease w_{Ap} results in syndrome w_{qr} and w_{qr} is the top-n feature of w_{Ap}. Similarly, $d_{pr}^{RF} = 1/2^{n-1}$ if syndrome w_{Ap} results from disease w_{qr} and w_{Ap} is the top-n feature of w_{qr}. The conditional probability of the i-th web pages with respect to aspect $s_{A,2}$ and query q is defined as

$$Sim_{axiom}(A_i, q) = \frac{Axiom(A_i, q)}{\sum_i Axiom(A_i, q)}.$$

3 Evaluation

To evaluate the proposed approach, a medical web search system was constructed. The web pages were collected from several Websites and totally 2322 web pages for medical domain and 8133 web pages for contrastive domain were collected.

On the other hand, the training and test queries for training and evaluating the system performance were also collected. Forty users, who do not take part in the system development, were asked to provide a set of queries given the collected web pages. After post-processing, the duplicate queries and the queries out of the medical domain are removed. Finally, 3207 test queries mixed Chinese with English words using natural language were obtained.

3.1 Keyword-Based VSM Approach: A baseline system for comparison

In recent years, most of the information retrieval approaches were based on the Vector-Space Model (VSM). Assuming that the query is denoted as a vector $q = (q_1, q_2, \ldots, q_n)$ and the web page is

represented as a vector $A=(a_1,a_2,...,a_n)$. The Tf-idf measure is employed and the similarity can be measured by the cosine function defined as follows.

$$Sim_{keyword-based}(A_i,q)=\cos(a,q)=\frac{\sum_{i=1}^{n}a_iq_i}{\sqrt{\sum_{i=1}^{n}a_i^{\,2}}\times\sqrt{\sum_{i=1}^{n}q_i^{\,2}}} \quad (8)$$

where $\|a\|=1$. This approach for key term expansion based on synonym set is also adopted in the baseline system. The results and discussions are described in the following sections.

3.2 Weight determination using 11-avgP score

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

$$Sim(A_i,q)=(1-\alpha)Sim_{relation}(A_i,q)+\alpha\times Sim_{axiom}(A_i,q) \quad (9)$$

This experiment is conducted on the estimation of the combination weights for each model. The results are shown in Figure 5. The performance measure called 11-AvgP [Eichmann and Srinivasan 1998] was used to summarize the precision and recall rates. The best 11-AvgP score will be obtained when the weight $\alpha=0.428$.

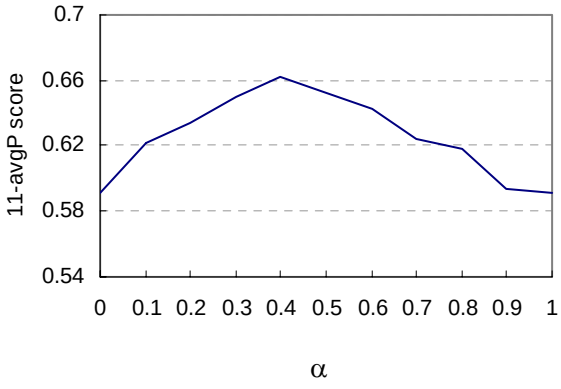

Figure 5 The 11-avgP score with different values of α

3.3 Evaluation on 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 is set to 1 for one inference module and the other is set to 0 in each evaluation. For comparison, the keyword-based VSM approach and the ontology-based system are also evaluated and shown in Figure 6. The precision and recall rates are used as the evaluation measures. And the ontology based approach means the combination of concept inference and axiom inference described in the section 3.2.

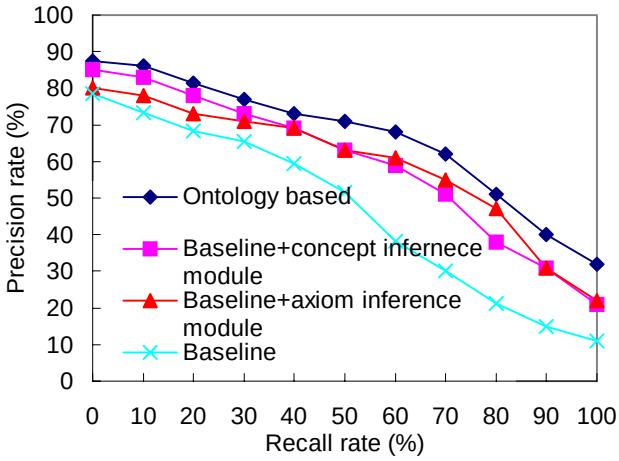

Figure 6 The precision rates and recall rates of the proposed method and the baseline system

4 Conclusion

This paper has presented an approach to automated ontology alignment and domain ontology extraction from two knowledge bases. In this approach, a bilingual ontology is developed using a corpus-based statistical approach from two well established language-dependent knowledge bases, WordNet and HowNet. A domain-dependent ontology is further extracted from the universal ontology using the island-driven algorithm and a domain corpus. In addition, domain-specific terms and axioms are also added into the domain ontology. We have applied the domain-specific ontology to the web page search in medical domain. The experimental results show that the proposed approach outperformed the keyword-based and synonym expansion approaches.

References

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

Christiane Fellbaum, 1998 WordNet an electronic Lexical Database, The MIT Press 1998. pp307-308

Fensel, D., Bussler, C., Ding, Y., Kartseva1, V., Klein, M., Korotkiy, M., Omelayenko, B. and Siebes R. 2002 Semantic Web Application Areas, the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB02).

M. Missikoff,, R. Navigli, and P. Velardi. 2002 Integrated approach to Web ontology learning and engineering, Computer , Volume: 35 Issue: 11 . 60 –63

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

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

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

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

H. Kargupta, I. Hamzaoglu, and B. Stafford. 1997. Distributed data mining using an agent based architecture. In Proceedings of Knowledge Discovery and Data Mining, 211-214.

U. Y. Nahm and R. J. Mooney. 2000. A mutually beneficial integration of data mining and information extraction. In Proceeding of the AAAI-00.

H. Ahonen, O. Heinoen, M. Klemettinen, and A. Verkamo. 1998. Applying data mining techniques for descriptive phrase extraction in digital document collections. In Advence in Digital Libraries.

E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In proceding of IJCAI-99, 668-673.

Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald, and X. Liu. 1999. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, 14(4):32-43.

W. W. Cohen. 1998. A web-based information system that reasons with structured collocations of text. In Proceedings of 2nd Agent'98

M. Junker, M. Sintek, and M. Rinck. 1999 Learning for text categorization and information extraction with ilp. In Proceedings of the Workshop on Learning Language in Logic, Bled, Slovenia

S. Oyama, T. Kokubo, and T. Ishida. 2004 Domain-Specific Web Search with Keyword Spice. IEEE Transactions on Knowledge and Data Engineering, Vol 16,NO. 1, 17-27.

Sankar K. Pal, Varun Talwar, and Pabitra Mitra. 2002. Web Minging in Soft Computing Framework: Relevance, State of the Art and Future Directions. IEEE Transactions on Neural Networks, Vol. 13, NO. 5.

T. Hofmann. 1999. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In Proceedings of 16th IJCAI, 682-687,

P. Martin and P. Eklund. 2000. Knowledge Indexation and Retrieval and the Word Wide Web. IEEE Intelligent Systems, special issue "Knowledge Management and Knowledge Distribution over the Internet"

H. Chen, A. M. Lally, B. Zhu, and M. Chau. ,2003 HelpfulMed: Intelligent Searching for Medical Information over the internet. Journal od the American Society for Information Science and Technology, 54(7):683-694.

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