國立交通大學

資訊科學與工程研究所

碩士論文

以知識本體為基礎之醫藥問答系統

Ontology-based Question Answering in Medicine

研究生：黃立泓

指導教授：梁　婷 教授

中華民國 九十五 年 六 月

以知識本體為基礎之醫藥問答系統

Ontology-based Question Answering in Medicine

研究生：黃立泓　　Student: Li-Hong Huang

指導教授：梁婷　　Advisor: Tyne Liang

國立交通大學

資訊科學與工程研究所

碩士論文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

以知識本體為基礎之醫藥問答系統

研究生：黃立泓　　　　　指導教授：梁婷

國立交通大學資訊科學與工程研究所

摘　要

自動醫藥問答在處理問題時牽涉到知識本體的運用、問題分析與資訊擷取。近年來 Unified Medical Language System (UMLS)大多被使用在醫藥領域上的知識查詢擴張，不同於以往專注在 UMLS 的查詢擴張研究，我們使用 UMLS 中概念的想法來萃取訓練語料中所產生的 Concept-Verb-Concept 樣本(CVC 樣本)，進而改善答案文本的排名。在問題分析方面，我們藉由 Naïve-Bayes 分類器將問題分成四個類別，依序為:診斷、治療、病因和定義。問題類別在擷取相關答案文本上被視為一個重要的基準，並透過查詢擴張來增加答案文本的召回率，結合 TF-IDF 和 CVC 樣本的權重衡量將答案文本排名。從資料量為 203 個問題的實驗結果顯示，所提出的問答系統平均 Mean Reciprocal Rank (MRR)值為 0.63。

關鍵詞：問答系統、知識本體、醫藥

Ontology-based Question Answering in Medicine

Student: Li-Hong Huang　　Advisor: Tyne Liang

Institute of Computer Science and Engineering
National Chiao-Tung University

Abstract

Automatic medical question answering involves the utilization of domain ontology, question analysis and information retrieval to process the medical question. Recently, Unified Medical Language System (UMLS) has been commonly utilized as the domain knowledge for medical query expansion. Unlike most previous researches focusing on UMLS as the domain expansion, we use the concepts in UMLS to extract Concept-Verb-Concept patterns (CVC patterns) from training corpus so as to improve the rank of answer texts. The proposed question analysis is to classify the questions into four categories based on Naïve-Bayes classifier, namely: diagnosis, therapy, etiology, and definition. The category is a basis to retrieve the relevant answer texts from PubMed and query expansion is used to increase the recall for document retrieval. The answer texts are ranked by combining the weight of TF-IDF and CVC patterns. The experimental result with 203 questions shows that the proposed QA can yield 0.63 Mean Reciprocal Rank (MRR).

Keyword: Question answering, Ontology, Medicine

Acknowledgement

學生能夠順利完成碩士論文，首先感謝梁婷老師對學生的教導，使學生對資訊擷取與自然語言處理產生濃厚的興趣，在論文研究上，老師時常針對學生的問題提供寶貴的意見，讓學生受益良多，並且教導學生如何正確地從事論文研究。

此外，立泓感謝實驗室每一位學長姐對我的關懷與協助，當立泓遭遇到困難時能夠得到及時的幫助，並且感謝實驗室的同學：守益、傳堯、曉茹，在研究過程中我們互相扶持，互相勉勵。

最後，我要感謝我的家人與女朋友—子誼，不斷地鼓勵我、支持我，使立泓能順利地完成碩士論文，謝謝你們。

僅將這篇碩士論文獻給所有在我身旁關懷我的人。感謝你們。

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background

The famous search engine, Google[1], receives more than 200 million queries every day. Automatic question answering becomes one of killer applications associated with natural language techniques and information retrieval to deal with. So it is desirable for computer scientists to propose efficient QA systems to extract the answers automatically.

Question answering researches have become popular since TREC[2] (Text REtrieval Conference) 1999. In TREC QA track, the QA systems proposed by the participators try to find the answer for a set of given questions from the collected documents provided by TREC. During the last decade, some QA systems have been proposed, such as START[3] presented by Katz et al. (2003). START is a Web-based QA system for several general domains including geography, science, arts, entertainment, history, and culture.

Recently, some researchers (Zhang et al. 2004; Niu et al. 2004; Wu et al. 2005) consider that specific domain QA has great potential. Specific domain QA is presented by using domain ontology. For example, Wu et al. (2005) divide the QA system into the question part and the answer part. They use the ontology proposed by Yeh et al. (2004) to calculate the distance of keyword concepts in the question part

[1] Google http://www.google.com.tw
[2] TREC http://trec.nist.gov/
[3] START http://start.csail.mit.edu/

and the casual relations in the answer part in order to retrieve the possible answer passages. Niu et al. (2004) consider the ontology as the specific expansion, such as hypernym expansion. Zhang et al. (2004) tag the categories for the nouns in the question and documents by using the ontology. The authors use okapi function to measure the similarity of categories between the question and the documents to retrieve the answer passages from the documents. In fact, how to utilize the domain knowledge is the main difference between open domain QA and specific domain QA. We discuss this topic in next section.

1.2 Specific Domain QA and Open Domain QA

Open domain QA processing involves question processing, information retrieval and answer extraction (Niu et al. 2004; John et al. 2004). Question processing is to understand what the question is asked about. The main purpose is to identify the answer type of a question so as to spot the answer. For open domain QA, the answer type can be identified by the interrogative word only. However, the interrogative word is not sufficient to understand query intention for specific domain. Take the questions “Who invented the toothbrush?” in open domain and “Who is at the greatest risk for heat-related illness?” in specific domain as the examples. We consider the answer type for two questions as person name according to the interrogative word. But the answer type is not person name for specific domain question. The details of examples are showed in Table 1.

Table 1. Examples of open domain and specific domain

	Question	Answer
Open Domain	Who invented the toothbrush?	William Addis
Specific Domain	Who is at the greatest risk for heat-related illness?	Infants and children up to four years of age, people 65 years of age and older …

The information retrieval module is to retrieve the relevant documents for the inputted question. In open domain QA, most of the questions are factoid questions, such as person, place, time, place or object. These questions are data-driven because their answers are always single. However, the domain knowledge is required for specific domain QA to understand the question and to consider whether the retrieved documents are relevant or not. So the specific domain questions are recognized as the knowledge-driven.

The answer extraction is to spot the answers from the relevant documents according to the information provided by the component of question processing. The strategy to locate the answers is calculating the similarity between the given question and the documents or passages. For example, the syntactic structure and named-entity are considered to spot the possible answers and the answers are ranked by the similarity score. In open domain QA, there is an explicit answer for each question, such as date, person name, or place name. But in specific domain QA, most of the specific domain questions are to concern the explanation.

1.3 Motivation

In this thesis, we concern the need to propose an efficient method for answering medical questions generated from people. The medical FAQs from the Web are the main data set for us to develop the medical QA because the questions of FAQs are generated by people and the answers of FAQs are provided by domain experts. They are good materials to propose specific domain QA.

For the medical QA, we use UMLS[4] (Unified Medical Language System) as knowledge base and PubMed[5] as the document source to deal with medical questions. First, the medical FAQs and medical literatures are collected from the Web. For the medical literatures, we extract the syntactic pattern as the form of NP-Verb-NP patterns. After concept identification for the noun phrases by using UMLS, the NP-Verb-NP patterns are transformed into Concept-Verb-Concept patterns. For the medical FAQs, the questions are used to train the question classifier. We also use the ontology to expand the query presented in (Hersh et al. 2000). When the question is inputted, the question is analyzed and the syntactic pattern with concept is identified by UMLS. The relevant texts which the answer may contain in are retrieved and ranked by scoring the weight of concept patterns and the weight of keywords.

There are three indicators for evaluating our method. The first indicator which we use to evaluate the performance of the method is the mean reciprocal rank (MRR). If the k-th abstract returned by the search engine contains the answer, the value of reciprocal rank is 1/k. The second indicator is human effort (HE). It is defined as the

[4] UMLS http://www.nlm.nih.gov/research/umls/

[5] PubMed http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

user finds the answer in the least rank of passages returned by the system. The third indicator is recall at top five passages returned. We take 203 questions from FAQs to evaluate the method. The experimental results show that there are 0.63 in MRR, 2.55 in human effort and 80% recall at top five passages for our proposed method.

The rest of the thesis is organized as follows. The related work is surveyed in Chapter 2. Medical question answering is described in Chapter 3. The evaluation and analysis are showed in Chapter 4. The conclusion and future work are given in Chapter 5.

Chapter 2. Related Work

Many researches (Zhang et al 2004; Niu et al. 2004; Soo et al. 2004; Wu et al. 2005) related to specific domain QA have been reported during the last decade. The specific domain QA is usually considered into four steps: the utilization of domain ontology, question processing, document retrieval and answer processing. For the domain ontology, it is the knowledge source for specific QA. The strategy to extract the relevant information by using domain ontology is the most important. Zhang et al. (2004) use the concepts of ontology to tag the question and the documents in order to measure the similarity between the question and the documents. Niu et al. (2004) consider the ontology as the keyword expansion for the question in order to gain more information. But to combine the ontology and the Web resources is another trend for specific domain QA. The system proposed by Soo et al. (2004) can integrate the biological literatures from the Web into the ontology automatically. Wu et al. (2005) use the medical FAQ from the Web as the data source to propose the medical QA. In the thesis, we consider how to utilize the concepts of ontology and the medical recourses, i.e. medical FAQ and literatures, to propose the method to deal with the medical questions in question processing and document retrieval.

For question processing, most specific domain QA adopts question classification as the essential component to deal with the given questions because there are different strategies to process the questions. Researches classify the question by identifying the format of answers, such as Yes/No format (Wu et al 2005), Description format (Wu et al 2005; Zhang et al. 2004) and NE format (Zhang et al. 2004). Except the question classification, how to extract the information from the given question by using the ontology is an important factor for the performance of document retrieval. In our study, the concept information and the syntactic relation from the given question are concerned in order to make document retrieval work efficiently. But the concept ambiguity is occurred during the processing. Navigli et al. (2005) provide a knowledge-based approach to do word sense disambiguation. They propose structural semantic interconnections algorithm (SSI) to construct the related senses as the form of network. The relations in the network are defined as the form in WordNet. In our study, the frequency of co-occurrence in UMLS is used to identify the concept.

For document retrieval, Zhang et al. (2004) use the okapi function to score the question concepts and keywords for retrieving the documents. Niu et al. (2004) show that the role information in the given question and documents is an important clue to match the relevant documents. On the other hand, query expansion will increase the performance for document retrieval. But 70% errors in handling QA are attributed to question classification, keyword selection, and query expansion as Moldovan et al. (2002) mentions. It is important that how to make query expansion efficient in document retrieval. Wang et al. (2004) propose Web-based unsupervised learning to transform the question term. They collect the QA pairs from Quiz-Zone as training corpus and align the question terms and bigrams of answer passages returned by Google. The question terms include the question keywords and the question patterns extracted by rules. The authors calculate the value of logarithmic likelihood ratio (LLR) between question terms and bigrams of answer passages and choose the top-rank bigram for each question term. These bigrams are recognized as the transformations for the question terms. The experiments indicate 0.69 MRR for the search engine to retrieve efficiently according to the keywords and expansions. But there is still sparseness problem for this method. On the other hand, Hersh et al. (2000)

use the relations in the UMLS Metathesaurus to expand the query. The relations include synonym, parent relation, child relation, and others. The authors consider the hierarchical relations as the important clue to increase the performance in document retrieval.

Zhang et al. (2004) constructed a specific domain QA via the ontology which connects the concepts by links like network. The ontology which they use is The Canadian thesaurus of Construction Science and Technology. The authors tag the parent category for the terms in the documents collected from Web according to the concepts of ontology. For an inputted question, the system will extract the headword by identifying the first head noun in the question and tag the category for the identifying word. The given questions are classified into four classes: definition, named entity, category and keyword type. The authors use the Okapi function to measure the weight of keywords for the passages in the documents and match the categories between the passage and the given question by counting the categories in common. Finally, they combine the weight of keywords and the number of categories by using linear function to rank the candidate passages. The result shows that the MRR value is 0.6545 and the improvement in the performance is 7.19%. It will decrease the recall in IR module because query expansion is not adopted for this system.

The Medical Question Answering system (MQA) is presented in (Niu et al. 2004) in which the PICO format presented by Straus et al. (2000) is used to deal with the given medical question and WordNet[6] and UMLS are used as the knowledge bases. WordNet is used to get the common keyword expansion and UMLS is used to get the

[6] WordNet http://wordnet.princeton.edu/

specific keyword expansion. The roles of PICO format are extracted from the questions. The authors match the roles between the question and the medical documents in order to spot the possible answer. The PICO format is considered as the important information in medical texts because the roles in the format construct the meaning of the text.

Additionally, the ontology is the most important resource in most of specific domain QA system. Soo et al. (2004) propose an agent to extract the knowledge from biological literatures. The authors integrate the knowledge resources, such as WordNet, MeSH[7](Medical Subject Heading), and GO[8](Gene Ontology), and develop the system to process the semantic annotation for the biological literatures automatically in order to encourage the domain knowledge in the ontology. For the inputted query, the system will infer the answers by using pattern matching and sentence parsing. The evaluation indicates that there are 85.2% in recall and 74.2% in precision. It improves recall from 48.1% to 85.2% and precision from 61.9% to 74.2% for the ontology-based knowledge extraction compared with the keyword-based search.

The FAQ are also considered as good materials to construct the medical QA because the answers are maintained by the domain experts. Wu et al. (2005) use the FAQ retrieval system to collect the medical FAQ pairs and adopt the medical ontology proposed by Yeh et al. (2004). The structure of ontology is based on WordNet and HowNet[9]. The authors consider the topic into two parts: question part and answer part. Three aspects are investigated separately for the question part, i.e. the question

[7] Medical Subject Heading (MeSH) http://www.nlm.nih.gov/mesh/meshhome.html
[8] Gene Ontology (GO) http://www.geneontology.org/
[9] HowNet http://www.keenage.com/

stem for the interrogative word, the distance of keyword concept in ontology, and the vector space representation between the FAQ questions and the inputted query. Two aspects are investigated separately for the answer part, i.e. the relations and the paragraph cluster. The relations in the ontology are identified for the answers of FAQ. They paragraph and cluster the answers of FAQ by using latent semantic analysis (LSA) and K-means algorithm in the paragraph cluster. The authors calculate the similarity for each aspect by conditional probabilistic function and combine those values by probabilistic mixture model. The EM algorithm is employed to optimize the mixing weights in the model. The answer formats are classified into three groups. The Set type means that the answer for the given question is enumerated. The Description type is the explanation for the given question. And the Boolean type is Yes/No question. The experimental results show that the Boolean type is 0.6643, the Set type is 0.6732, and the Description type is 0.6327 for the metric of 11-AvgP.

For answering definitional questions, Hovy et al. (2001) use WordNet to assist the QA to deal with them. In recent years, Xu et al. (2004) consider the linguistic features as the important clues to extract the definitions from the documents. With the growth of Web, Hildebrandt et al. (2004) use the surface patterns to collect the definitions from Web and integrate the definitions into knowledge database in order to answer this type of questions. In the thesis, we use the definition database from UMLS to answer the definitional question. If the definition is not found in it, the online dictionary is queried to answer the question and expand the definition database at the same time.

Xu et al. (2004) use the linguistic features to extract the definitional information from the documents. They take five types of ranked features to handle the definitional questions in the following order: appositives, copulas, structured patterns, relations, and propositions and establish the question profile for definitions from many sources, such as WordNet glossaries, Merriam-Webster dictionary, Columbia encyclopedia and Google. They calculate the similarity of given question according to the question profile. The top ten features are selected for the given question by using the similarity score. The five ranked features and the top ten features are used to extract the definitions from the documents. The experiment shows 0.555 for F-score in performance.

On the other hand, Hildebrandt et al. (2004) want to answer definitional questions by using multiple knowledge sources on the Web. They collect the definitional answers by using surface patterns and normalize them as the form of database. If the answer can’t be found in the collected data, the authors will process the question into the string and query the online dictionary or document retriever. In our study, we will detect the definitional question first by using simple patterns and use UMLS ontology to answer this type of questions. We will convert the question into a single noun phrase and retrieve the definition from Web dictionary if the definition is not found in UMLS.

For the relevant work on specific domain QA, we focus on the problem in converting the given question into the syntactic relations with concept identification by using UMLS and integrating the medical literatures from PubMed as the document source to match the relevant passages or documents by mixing the weight of TF-IDF score and Concept-Verb-Concept score.

Chapter 3. The Proposed QA Method

3.1 Data Collection

We collect 910 FAQs from some medical Web, such as FDA[10], NCI[11], WHO[12], HHS[13], and CDC[14]. Table 2 shows the sources of QA pairs in detail. Most of the collected questions are not the factoid questions according to their answer type. The average length for each question is 9.5 words and the average length for each answer is 130.1 words. Figure 1 shows that there are 83.3% for the interrogative words of "what" and "how" in the collected data. Figure 2 shows the distribution of semantic categorizations in the collected FAQs. On the other hand, we also use 400 medical terms as the keywords in UMLS to query PubMed and collect 8,729 medical abstracts for training materials of NP-Verb-NP patterns in order to extract Concept-Verb-Concept patterns by using the concepts in UMLS.

Table 2. Data sources

	Number of QA pair	Average Length of Q	Average Length of A
FDA	20	11.6	119.2
NCI	174	8.7	105.7
WHO	22	7.2	139.2
HHS	50	11.2	166.4
CDC	644	8.9	120.4
ALL	910	9.5	130.2

[10] U.S. Food and Drug Administration (FDA) http://www.fda.gov/
[11] National Cancer Institute (NCI) http://www.cancer.gov/
[12] World Health Organization (WHO) http://www.cancer.gov/
[13] United States Department of Health and Human Services (HHS) http://www.hhs.gov/
[14] Centers for Disease Control and Prevention (CDC) http://www.cdc.gov/

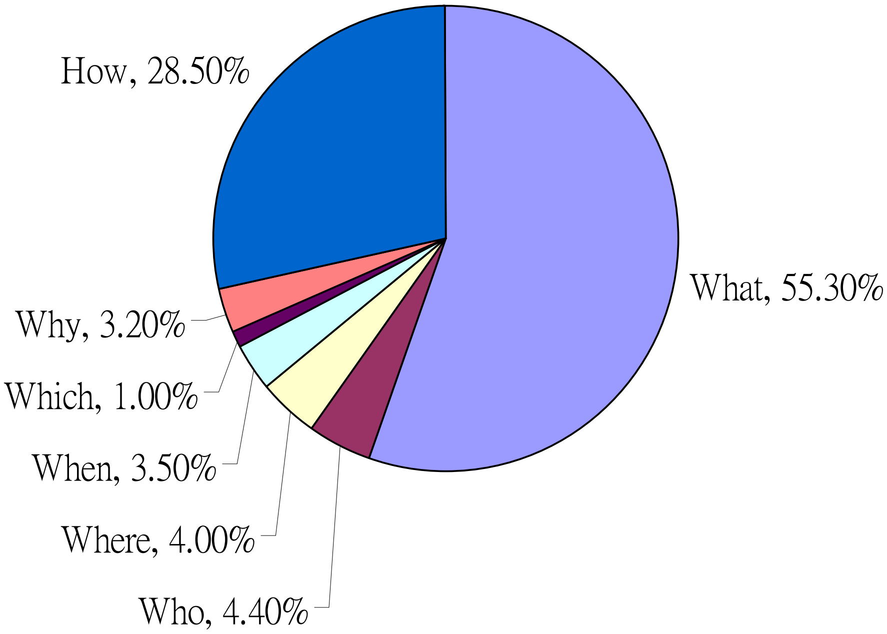

Figure 1. Question word

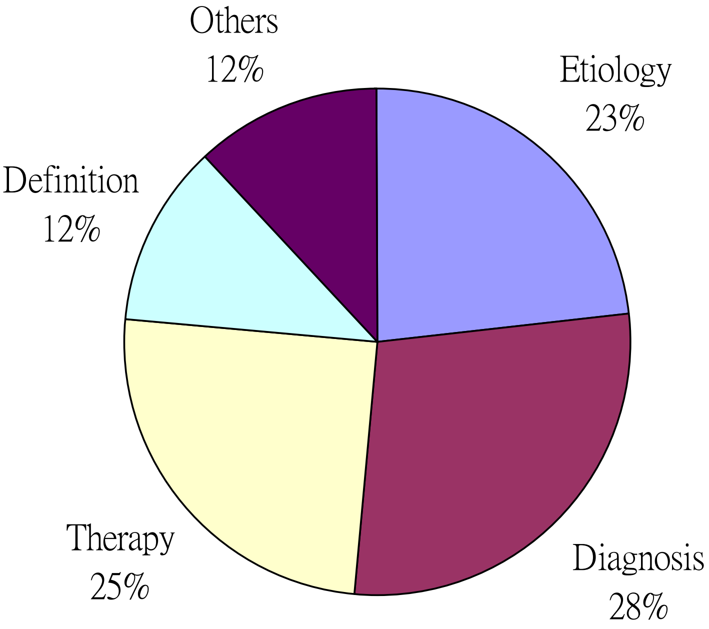

Figure 2. Semantic categorization

3.2 Tagging and Parsing

We use the English Part-of-Speech tagger[15] which is proposed by NLM[16]. The tool assigns the POS tags and phrase tags to the inputted texts, such as questions or medical texts. This tagger is good for medical texts because it includes over 66,000 medical terms in the dictionary. The full parser we use in the thesis is MINIPAR[17] so as to get the dependency structure while analyzing the definitional questions.

3.3 QA Processing

The proposed QA processing as shown in Figure 3 can be divided into several components. First, the definitional step will detect the given question whether the question is definitional type or not. If the question is definitional type, the definitional strategy will be involved to process the question. If the question is the other types, we use a Naïve-Bayes classifier to classify the questions into proper types and identify the concept of noun phrases by UMLS in the NP-Verb-NP pattern extracted from the question. The question type and Concept-Verb-Concept pattern (CVC patterns) are identified in question processing in order to calculate the weight of answer texts returned from search engine in information retrieval phase. On the other hand, we use ontology-based expansion proposed by Hersh et al. (2000) to expand the query in order to increase the recall for retrieving the relevant data. Finally, we measure the weight of the returned texts by TF-IDF and Concept-Verb-Concept, and re-rank the texts as the result.

[15] Part-of-Speech Tagger http://tamas.nlm.nih.gov/tagger.html
[16] National Library of Medicine (NLM) http://www.nlm.nih.gov/
[17] MINIPAR http://www.cs.ualberta.ca/~lindek/minipar.htm

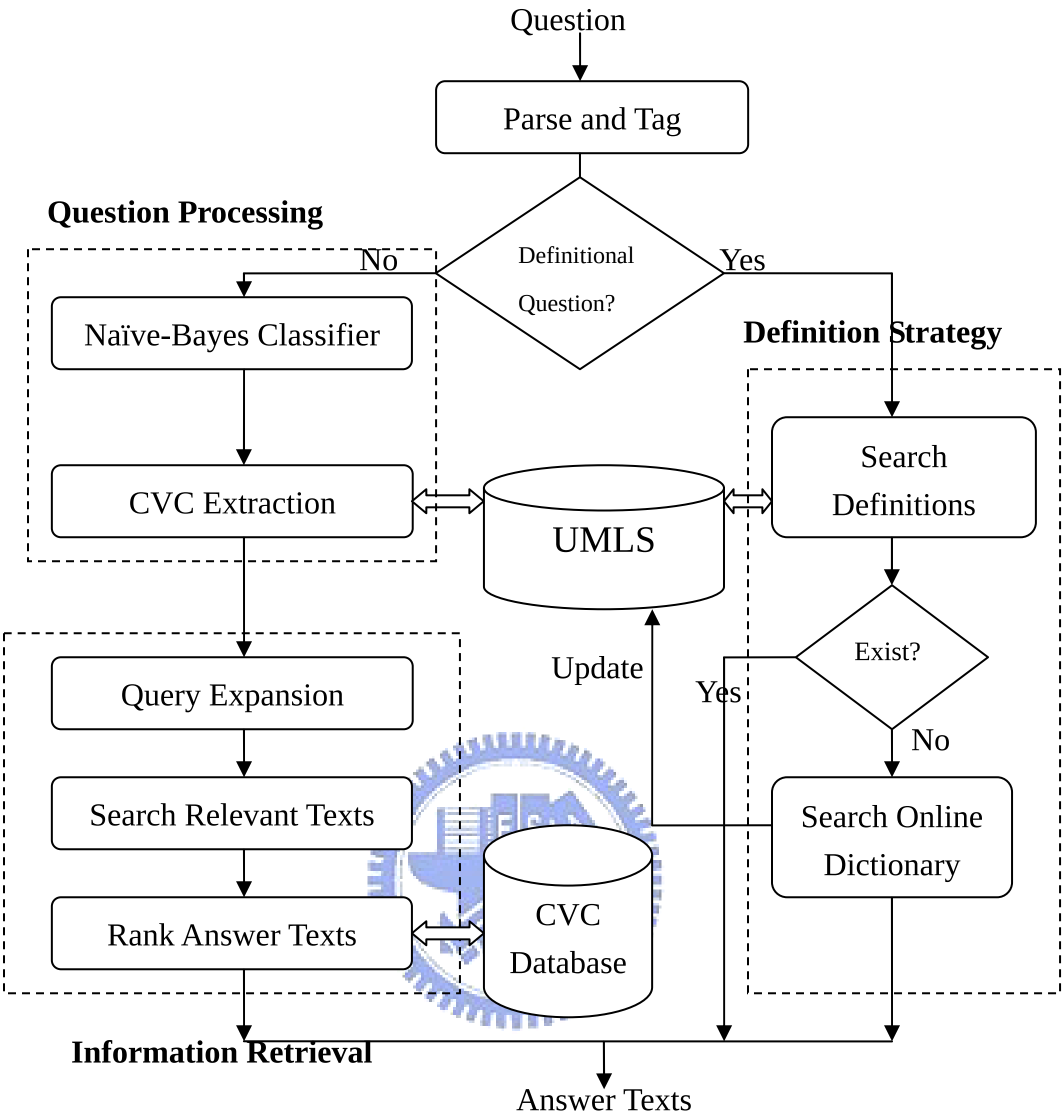

Figure 3. Flowchart of QA processing

3.4 Rule-based Approach for Definitional Question Identification

The main idea to approach definitional questions is from (Hildebrandt et al. 2004) which collect the definitions from Web and integrate the definitions into knowledge database. The knowledge sources we apply are UMLS. And we update the definition database by retrieving the latest definition from the online dictionary (Merriam-Wesbster[18]).

3.4.1 Features of Definitional Question

We use MINIPAR to parse the definitional questions. There are 108 definitional questions which have been classified manually in 910 pairs of the collected FAQs. We parse these questions and analyze the sentence structure. 88% of definitional questions are parsed as the following two structures.

Structure 1: (What OR Who) + be + ((Term$_1$) (Term$_2$) (Term$_3$)…headword)

Example: “What is the anthrax vaccine?”

⇨ What + be + ((the) (anthrax) vaccine)

Structure 2: (What OR Who) + be + ((Term$_1$ (Term$_2$ (Term$_3$ (…)))) headword)

Example: “What is West Nile virus?”

⇨ What + be + ((West Nile (West)) virus)

Table 3. The coverage rate for each structure

	Number of Questions	Coverage Rate
Structure 1	48	44.4%
Structure 2	47	43.5%
Structure 1 + Structure 2	95	88%

The headword is the root for the parsing tree of noun phrase. In structure 1, the

[18] Merriam-Wesbster [http://www.m-w.com/]

headword connects the other terms parallel in the parsing tree, e.g. “What is the anthrax vaccine?” In structure 2, the headword connects the other terms hierarchically in the parsing tree, e.g. “What is West Nile virus?” The parser will recognize the noun phrase as the subject of sentence in two structures. And then we can take the noun phrase to search the definitions in UMLS.

The rules used to recognize definitional questions are listed as follows:

(i). The length of POS sequence is less and equal than four

(ii). [“What or Who” + “be” + NP], the question structure is identified as structure 1 or structure 2

(iii). The question contains only one NP

(iv). There are no prepositions in NP

3.4.2 Test Results on Definitional Questions

In the experiment, we take 40 definitional questions from TREC-9 to evaluate the definitional rules. The experimental results show that 36 questions are detected by these rules. The accuracy rate is 90% in the test data. The error rate for detecting definitional questions is about 10%. The errors are caused by the wrong parsing tree or tags.

Table 4. Test results on definitional questions

	Developing	Testing
Number of Questions	95	40
Number of Correct Type	95	36
Accuracy Rate	100%	90%

For non-definitional questions, we use a Naïve-Bayes classifier to determine the question type. In next section, the features for the classifier are discussed and evaluated in the metric of recall and precision.

3.5 Naïve-Bayes Classifier for Other Type Questions

A Naïve-Bayes classifier is used to classify the non-definitional questions into the pre-defined types, namely: diagnosis, therapy and etiology. We collect 8,729 medical documents which have been classified from PubMed as the training data. The documents returned from PubMed are segmented as the form of n-gram except trigram. We calculate the probability of n-grams and filter out the n-grams which contain the stop words or medical proper nouns in UMLS. The n-grams are clustered into 18 groups by a typical K-means algorithm. For the collected questions, we extract POS sequence from the classified questions and analyze POS sequence as the feature for our classifier.

We follow the Bayesian theorem to train the question classifier by the features of n-grams and POS sequence. The probabilistic model is described as follows.

$$\mathrm{Pr}ob_c = \mathrm{argmax}_c\, P(C)\prod_{k=1}^{3} P(F_i \mid C) \qquad (1)$$

C = {diagnosis, therapy, etiology}

F_i = {unigram, bigram, POS sequence}

The probabilistic model is used to calculate the values for each question type. We take the highest value and assign the type for the question. In the testing phase, we take 453 questions randomly from the rest FAQs. There are 85% precision and 86% recall for diagnosis, 84% precision and 94% recall for therapy and 82% precision and 88% recall for etiology. There are some examples about the classification in Table 5 through Table 7.

Table 5. Testing results on non-definitional questions

Type	Diagnosis	Therapy	Etiology
System Classified	207	122	124
TP+FP	205	109	115
TP	176	102	101
Precision	85%	84%	82%
Recall	86%	94%	88%

Table 6. Frequent n-grams for each type

Type	Unigram	Bigram	POS Sequence
Diagnosis	symptom, case, diagnosis, syndrome	diagnosis of, case of, symptom of	np vp np pp, np vp np pp pp, vp np vp pp
Therapy	treatment, therapy, use, treat	treatment of, treat with, be treat	np vp np, vp np vp, np vp np pp
Etiology	prevent, cause, involve	to prevent, cause of, prevent the	vp np vp np, vp np vp pp pp, np vp np

Table 7.　Examples of question classifier

Question	Original Type	Classifier
How is Japanese encephalitis treated?	Therapy	Therapy
What are the symptoms of diabetes?	Diagnosis	Diagnosis
What is the treatment for diabetes?	Therapy	Therapy
What causes HFMD?	Etiology	Etiology
How is OPC diagnosed?	Diagnosis	Diagnosis
What are the risk factors for hepatitis B?	Diagnosis	Diagnosis
How is asthma normally treated?	Therapy	Therapy
What drugs are used to treat chronic hepatitis B?	Therapy	Therapy

3.6 Concept Identification

After question classification, we extract the NP-Verb-NP pattern from the given question. Concept identification is presented to distinguish the concepts for each medical phrase in the question in order to transform the NP-Verb-NP pattern into Concept-Verb-Concept pattern. UMLS is the multi-node structure which a string may appear in different path for the hierarchical tree. It is necessary to do concept disambiguation in order to assign the most possible concept to the noun phrases in the question. The method is that we use the co-occurrence information in UMLS to calculate the weight among the noun phrases which are extracted from the question. The concept probabilistic function is designed as equation (3).

After the calculation of this probabilistic function, all concepts for the noun phrases are calculated with the probabilistic value by using UMLS. Then we use the association function to measure the concepts which are the most possible to be associated in the sentence. The association function for these concepts is defined by equation (2). The identification steps are summarized as following.

$$Association(X_r, Y_h) = \Pr ob(X_r - > Y_h) * \Pr ob(Y_h - > X_r) \tag{2}$$

$$\Pr ob(X_r - > Y_h) = \frac{freq(X_r, Y_h)}{freq(X_r, *)} \tag{3}$$

$X_r \in \{X_1, X_2\ldots, X_i\}, Y_h \in \{Y_1, Y_2\ldots, Y_j\}$

freq(X_r, *): the co-occurrence which contains concept X_r

freq(X_r, Y_h): the co-occurrence for concept X_r and Y_h

We use the Algorithm of Concept Identification to identify the concepts of noun phrases according to UMLS. NP-Verb-NP pattern is formed as the tuple of [Concept_A, Verb, Concept_B].

Algorithm for Concept Identification

```
If the question contains only one noun phrase
    Then we get all concepts for the noun phrase from UMLS
Otherwise
    (i).   Identify all concepts for noun phrases
    (ii).  Calculate the probability for all concepts of the noun phrases according
           to the co-occurrence in UMLS
    (iii). Calculate the associative value to choose the most possible concept by
           equation (3) and assign it to the noun phrase
```

We consider the question which contains the terms, “AIDS” and “HIV”. First, the concepts for “AIDS” and “HIV” are identified by using UMLS. The probability for all concepts is calculated by equation (2). We use equation (3) to calculate the associative degree and choose the concept with the top value to identify the noun phrase. There is an example described as follows. We consider that there are three concepts (C_1, C_2, and C_3) for “AIDS” and two concepts (C_4 and C_5) for “HIV”.

Table 8. All concepts for each term

Term	Concepts
AIDS	C_1, C_2, C_3
HIV	C_4,C_5

Table 9. Co-occurrence information for concept identification

Concept_A	Concept_B	Frequency
C_1	C_4	3
C_1	C_5	4
C_1	C_9	1
C_2	C_3	2
C_3	C_4	7
C_3	C_7	8
C_4	C_1	3
C_4	C_3	7
C_5	C_1	2
C_5	C_7	4

$$Association(C_1,C_4)=\Pr ob(C_1->C_4)*\Pr ob(C_4->C_1)=\frac{3}{3+4+1}*\frac{3}{3+7}=0.1125$$

$$Association(C_1,C_5)=\Pr ob(C_1->C_5)*\Pr ob(C_5->C_1)=\frac{4}{3+4+1}*\frac{2}{2+4}=0.1667$$

$$Association(C_2,C_4)=\Pr ob(C_2->C_4)*\Pr ob(C_4->C_2)=\frac{0}{2}*\frac{0}{3+7}=0$$

$$Association(C_2,C_5)=\Pr ob(C_2->C_5)*\Pr ob(C_5->C_2)=\frac{0}{2}*\frac{0}{2+4}=0$$

$$Association(C_3,C_4)=\Pr ob(C_3->C_4)*\Pr ob(C_4->C_3)=\frac{7}{7+8}*\frac{7}{3+7}=0.3267$$

$$Association(C_3,C_5)=\Pr ob(C_3->C_5)*\Pr ob(C_5->C_3)=\frac{0}{7+8}*\frac{0}{2+4}=0$$

Table 10. Result for concept identification

Term	Concept
AIDS	C_3
HIV	C_4

3.7 Training Phase for CVC Patterns

The main purpose of Concept-Verb-Concept patterns (CVC patterns) is used to score the answer texts in information retrieval. In the training phase, we use 400 medical terms as the keywords in UMLS to query the PubMed and collect 8,729 medical abstracts for training materials. The strategy is that noun phrase preceding and succeeding the verb are extracted in the medical abstracts. If the noun phrase is a pronoun, the noun phrase which is preceded or succeeded the pronoun is extracted instead of the pronoun. We combine noun phrases preceding and succeeding the verb as the format of NP-Verb-NP.

We extract NP-Verb-NP patterns from the training data and use the algorithm of concept identification to identify the concepts of noun phrases according to UMLS. And we collect Concept-Verb-Concept patterns in order to calculate the degree of the relation between Concept$_A$ and Concept$_B$. For the verb in CVC patterns, we use the synsets of verb in WordNet to cluster CVC patterns into 4,496 groups. The following tables show some results about NP-Verb-NP in Table 11. The degree function which we apply is described as follows.

$$Degree(CVC_t) = \frac{freq(C_A, Verb, C_B)}{freq(C_A, Verb) + freq(Verb, C_B) - freq(C_A, Verb, C_B)} \quad (4)$$

freq(C$_A$,Verb) = the co-occurrence for (Concept$_A$,Verb)

freq(Verb,C$_B$) = the co-occurrence for (Verb,Concept$_B$)

freq(C$_A$,Verb,C$_B$) = the co-occurrence for (Concept$_A$,Verb,Concept$_B$)

Table 11. Examples of NP-Verb-NP patterns

NP_A	Verb	NP_B
mouse	ameliorate	antibody
antioxidant	need	the defense system
carduus	evaluate	puccinia
dna	produce	unique pattern
dna	isolate	carduus
twin	develop	brain

At run time, we use CVC pattern extracted from the given question to retrieve the relevant CVC patterns from the training results. For information retrieval, the relevant CVC patterns are used to score the answer texts returned by search engine.

3.8 Ontology-based Query Expansion

On the other hand, there is not much information provided from the given question. To expand the keywords in the question is necessary for QA. So we propose a method which the idea is from (Hersh et al. 2000) to expand the query. The authors use the synonyms and hierarchical relations in UMLS Metathesaurus to expand the terms in the query. The expanded strategy is described as follows:

For each medical term in query

(i). Add the synonym variants in UMLS to the query

(ii). Add its parent terms in UMLS to the query

(iii). Add its child terms in UMLS to the query

(iv). Add other relations defined in UMLS to the query

There is an example showed for query expansion. We consider that the terms in the question, Acute tubular necrosis, Aminoglycosides, AIDS , and HIV, as the medical terms for expanding.

Table 12. Ontology-based expansion

Synonym	Acute tubular necrosis	Acute, atn, failure, ischemic, kidney, lesion, lower, necrosis, nephron, nephropathy
Parent term	AIDS	Abnormal, agent, antibody, behavior, disease, hiv, htlv
Child term	Aminoglycosides	Aminoglycosides, Amikacin, Amikacin Sulfate, Butirosin Sulfate, Framycetin, Genticin, Gentamicins
Other relation	HIV	Adult, anxiety, assay, arthritis, blood, body

3.9 Retrieval Procedure for QA

The retrieval procedure in our method is that we use PubMed as the major information retrieval platform and Google as the minor platform. For PubMed, there are three aspects: etiology, diagnosis and therapy for us to retrieve the abstracts of medical literatures. Our question classifier will detect the question type for the inputted question and trigger PubMed to retrieve the relevant medical texts. For Google, if there is no relevant data in PubMed for the question, Google will be triggered to retrieve the snippets according to the keywords from the given question.

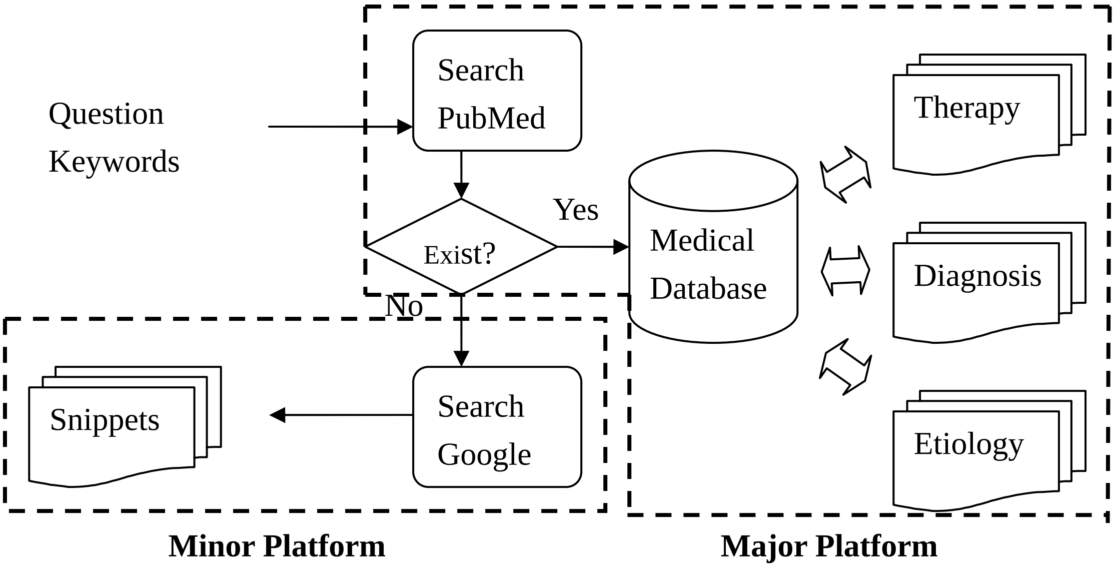

Figure 4. Retrieval procedure

3.10 Rank by TF-IDF and Concept-Verb-Concept

In the previous section, whether the data is returned by PubMed or Google, we measure the answer texts by using TF-IDF function. The question keywords and query expansions are used to calculate the weight for the answer texts. After the processing of TF-IDF, we get the initial rank for each answer text. The rank is considered as TF-IDF rank in the following processing.

$$\sum W_{i,j} = \sum (0.5 + \frac{0.5\, freq_{i,j}}{\max \; freq_{i,j}}) * \log \frac{N}{n_i} \qquad (5)$$

$freq_{i,j}$: the frequency of term i in the document j

N : the number of documents

$n_{i,j}$: the number of documents contained term i

Additionally, we will extract NP-Verb-NP patterns from the given question and identify the concepts of noun phrases in the patterns by using UMLS. The NP-Verb-NP patterns are considered as the form of Concept-Verb-Concept patterns (CVC patterns). The concepts and hypernym concepts in the CVC patterns are utilized to retrieve the relevant CVC patterns from the database collected in the training phase. For the answer texts returned from search engine, CVC patterns are also extracted and identified. We match the CVC patterns between the given question and the answer texts. The CVC rank is measured by scoring the degree of the CVC patterns checked in common between the question and the answer texts.

In order to optimize the rank, TF-IDF rank and CVC rank are mixed as the final rank. The ranking function is described as follows: $\text{Rank}_{\text{avg}} = (\text{Rank}_{\text{TF-IDF}} + \text{Rank}_{\text{Concept-Verb-Concept}})/2$.

Table 13. Final Rank

	TF-IDF Rank	CVC Rank	Mixed Rank	Final Rank
Text 7	2	2	2	1
Text 5	1	4	2.5	2
Text 2	3	3	3	3
…..	…..	…..	…..	…..

Chapter 4.　Experiments and Analysis

4.1 Experimental Setup

In this chapter, we evaluate the implement of our method proposed in previous chapter. We collect 910 pairs of questions and answers from medical Web, such as FDA, NCI, WHO, HHS, and CDC. There are 203 questions which are set aside from the collected FAQs for testing purpose.

Three indicators are used to measure the performance for our method. One is the mean reciprocal rank (MRR). If the k-th passage returned contains the answer's information, then the reciprocal rank of the passage is 1/k. The MRR is the average reciprocal rank of the questions in the test corpus. Another is the human effort (HE). The human effort is defined as the user finds the answer in the least rank of passages returned. The other is recall at top five texts returned. In next section, we will describe and analyze the experimental results.

4.2 Performance of Medical Question Answering

For the Concept-Verb-Concept (CVC) patterns, we take 8,729 medical abstracts from PubMed to extract the patterns with concept identification by using UMLS. There are 951,678 distinct patterns received from the training set. The details of training results are described in Table 14.

Table 14.　Training result for Concept-Verb-Concept patterns

Data source	Number of Abstracts	Distinct CVC Patterns	Clusters
PubMed	8,729	951,678	4,496

We want to evaluate each component about our method. The precision for question classifier is important because our strategy will retrieve the relevant documents according to the question type and the strategy for each question type is not the same. For the definitional question, we use rule-based method to detect it and assign the tag of “definition” to the question. For the other type, the classifier will assign the tag according to the probability of n-grams clustered by K-means algorithm and POS sequence for each type.

We divide the method into three components: Question Classifier (QC), Query Expansion (QE) and Concept-Verb-Concept scoring (CVC). We take 55 questions from testing corpus to evaluate each component. The contribution for each component can be seen in Table 15.

Table 15. MRR of each component

	MRR
QC+QE+CVC	0.63
QC+QE	0.58
QC+CVC	0.57

In Table 15, the improvement of MRR is about 0.06 for query expansion. It consists with other researches (Wang et al. 2004; Niu et al. 2004) in question answering. Query expansion will provide some important patterns to retrieve the more correct documents for the given question. The improvement of MRR is about 0.05 for CVC patterns. The main idea for CVC patterns is to extract the implicit medical information as the form of patterns by using UMLS. The results show that specific domain knowledge help QA improve the performance.

Table 16. MRR for each semantic categorization

	Number of Questions	MRR
Diagnosis	103	0.62
Therapy	45	0.67
Etiology	55	0.62

We also take 203 FAQ questions which are set aside from the collected FAQs to evaluate the method (QC+QE+CVC) according to the question type. The experimental results show in Table 16. There are 0.62 for Diagnosis, 0.67 for Therapy and 0.62 for Etiology in MRR. According to the experimental results, therapy type is more efficient than the other type. Because diagnostic and causal conditions are similar in many cases and the information from the given question is not sufficient for QA.

Additionally, we also classify the testing questions only by using the interrogative words, such as what, where, when, who, why, and how. The evaluation is designed to analyze the intention of question simply according to the interrogative words. The result is showed in Table 17. For the interrogative word, there is only 0.54 MRR for the “when” type. The medical literatures always contain few information for the “when” type. This will causes the MRR value decreased for the “When” type.

Table 17. MRR for the interrogative words

	What	When	Who	Where	Why	How
Number of Questions	78	8	13	11	5	88
MRR	0.63	0.54	0.65	0.64	0.66	0.64

For the factoid questions, the interrogative word in the question can be determined its answer type. Some medical questions are factoid by our observation. For example, consider the question “What is the mortality rate of SARS?” In order to evaluate our method for factoid questions, we take 25 medical questions rewritten from TREC-8 to evaluate the method. The results are described in Table 18.

Table 18. MRR for TREC-factoid questions

Number of Questions	MRR
25	0.62

For the module of information retrieval, PubMed and Google are the document sources for our method. We count the number for the document source which is PubMed or Google in our method. The MRR value for each search engine is also calculated. The results are showed in Table 19.

Table 19. Result for each document source

	PubMed	Google
Number of Questions	54	149
Percentage	27%	73%
MRR	0.53	0.66

Another indicator is human effort (HE). We record the top five answer texts in statistical method and calculate the average of human effort for each experiment. In the experimental setup for human effort, we consider the question type as the major class to evaluate the method. The experimental results of human effort are described in Table 20. People can find the answer passage at the top 2 or top 3 in the returned texts. In Figure 5, we evaluate all types by using the indicator of recall at the top five passages. There is 79% recall at top five texts. In Figure 6, the curves show the increasing rate of recall for each question type. There are 79% recall for diagnosis, 80% recall for therapy and 80% recall for etiology at top five texts returned.

Table 20. Human effort for each component

Rank	Rank Count			
	Diagnosis	Therapy	Etiology	All Types
1	48	24	27	99
2	19	9	6	34
3	9	3	6	18
4	3	0	2	5
5	3	0	3	6
No Answer	21	9	11	41
# of questions	103	45	55	203
HE per question	2.58	2.33	2.65	2.55

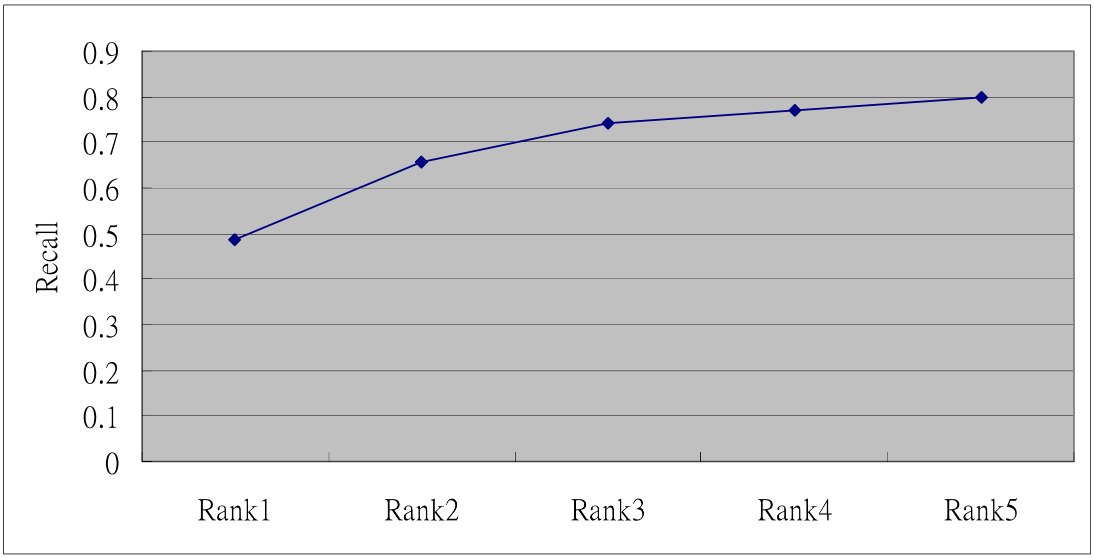

Figure 5. Recall for all types

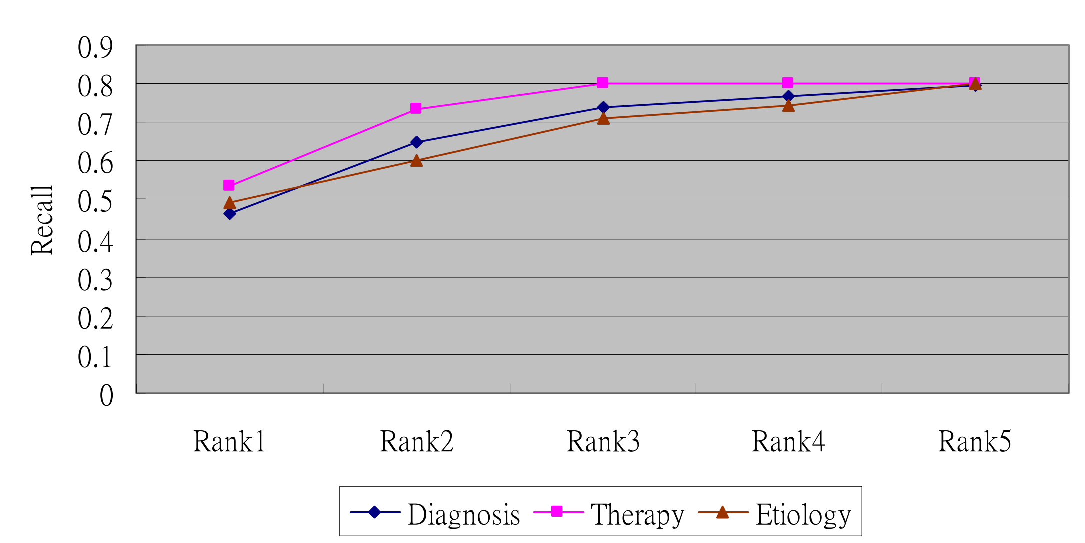

Figure 6. Recall for each type

By our observations for the experiments, there are some reasons caused to decrease the performance:

- Incorrect POS tagging.
- Assign the wrong category for the given question.
- Assigning the concept to each noun phrase that is not sufficient enough to explain the meanings

According to the experimental results, we find that the knowledge in the ontology indeed improve the performance of QA in specific domain. For CVC patterns, we also use hypernym concept as concept expansion for the CVC pattern from the given question to extract more medical implicit information. The experimental results show that the idea is positive for the performance. On other hand, query expansion by using the relations in UMLS works effectively to retrieve the more relevant documents from search engine. We integrate the medical resources from the Web into the question answering, such as online medical literature, UMLS resources. Natural language processing and information retrieval technique are the key points to integrate them for the users to get the answers from the huge amount of data.

Chapter 5. Conclusion and Future Work

5.1 Conclusion

In this thesis, we construct the medical domain QA by using the knowledge in UMLS. The hierarchical structure and the concept in the ontology provide more knowledge to expand the meanings in the question. CVC patterns can extract the implicit information contained in the question and the texts by using UMLS. At run time, our strategy is to use the rules to detect the definition question. If the question is definitional question, it will involve the strategy to process the question and retrieve the relevant definitions. If the question is the other type, our procedure is involved to deal with the give question according to its question type. First, we extract the features from the question as the input of Naïve-Bayes classifier and identify the concept of noun phrases by using UMLS. For query expansion, the keywords are automatically expanded by using the relations in UMLS and the answer texts are retrieved from the Web by using the keywords. Finally, we use TF-IDF function to measure the weight of keywords and score the weight of CVC patterns in each text. TF-IDF rank and CVC rank are mixed as the final rank for the re-ranking procedure.

The methodology for the medical QA is effective because it focuses on the following features:

- Tagging the concept for each noun phrase from NP-Verb-NP patterns provides a more general outlook for medical QA.
- Combine concepts, co-occurrence and hierarchical relations in UMLS to measure the questions or the answer texts by the Concept-Verb-Concept format.

- Combine the weight of keywords (TF-IDF) and the knowledge in UMLS (CVC patterns).

5.2 Future Work

There are some future directions for this topic. For answer spotting, how to summarize the appropriate passage from the answer texts automatically is a good study for specific domain QA. For the domain ontology, developing a medical ontology for medical QA provides more information to process the questions.

References

Ceusters, Werner, Barry Smith, Maarten Van Mol. “Using ontology in query answering systems: scenarios, requirements and challenges.” In Proceedings of *the 2nd CoLogNET-ElsNET Symposium,* Amsterdam, pp.5-15, 18 December 2003.

Duclaye, Florence, Francois Yvon, and Olivier Collin. “Learning Paraphrases to Improve a Question-Answering System.” In Workshop of *the European Chapter of the Association for the Computational Linguistics*, 2003

Hersh, William, Susan Price, and Larry Donohoe “Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus.” In Proceedings of *American Medical Informatics Association (AMIA) Annual Symp*, pp. 344–348, 2000.

Hildebrandt, Wesley, Boris Katz, and Jimmy Lin. “Answering Definition Questions Using Multiple Knowledge Sources.” In Proceedings of *the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004),* Boston, Massachusetts, pp.49-56, 2004.

Hovy, Eduard, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin-Yew Lin. “Question Answering in Webclopedia” In Proceedings of *the TREC-9 Question Answering Track* , pp.655-672, 2000.

Katz, Boris and Jimmy Lin. “Selectively Using Relations to Improve Precision in

Question Answering." In Proceedings of *11th Conference of the European Chapter of the Association for the Computational Linguistics*, pp.43-50, 2003.

Lin, Dekang. "Dependency-based Evaluation of MINIPAR." In *Workshop on the Evaluation of Parsing System*, May 1998.

McCray, Alexa T. "An upper-level ontology for the biomedical domain." Published online in Wiley InterScience.

Melamed, I. Dan. "A Word-toWord Model to Translational Equivalence." In Proceedings of *the 35st Annual Meeting of the Association for Computational Linguistics 1997*, pp.490-497, 1997

Moldovan, D., Pasca, M., Harabagiu S., and Surdeanu M. "Performance Issues and Error Analysis in an Open-domain Question Answering System." In Proceedings of *the 40th Annual Meeting of the Association for Computational Linguistic*, pp.33-40, 2002.

Navigli, Roberto, and Paola Velardi. "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 27, Issue 7, pp.1075 - 1086, July 2005.

Niu, Yun, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. "Answering Clinical Questions with Role Identification." In Proceedings of *the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pp.73-80, 2003.

Prager, John, Jennifer Chu-Carroll and Krzysztof Czuba. “Question Answering using Constraint Satisfaction: QA-by-Dossier-with-Constraints.” In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp.575-582, 2004.

Sang, Erik Tjong Kim, Gosse Bouma, and Maarten de Rijke. “Developing Offline Strategies for Answering Medical Question.” In Proceedings of *American Association for Artificial Intelligence*, 2005.

Shen, Dan, Geert-Jan M. Kruijff, and Dietrich Klakow. “Exploring Syntactic Relation Patterns for Question Answering.” In Proceedings of *International Joint Conference on Natural Language Processing 2005*. LNAI3651. pp.507-518, 2005.

Soo, Von-Wun, Hsiang-Yuan Yeh, Shih-Neng Lin, and Wen-Ching Chen**.** **“**Ontology-based Knowledge Extraction from Semantic Annotated Biological Literatures.” In Proceedings of *the Ninth Conference on Artificial Intelligence and Applications*, 2004.

Wang, Yi-Chia, Jain-Cheng Wu, Tyne Liang, and Jason S. Chang. “Using the Web as Corpus for Un-supervised Learning in Question Answering.” In Proceedings of *ROCLING 2004*, pp.191-198, 2004.

Wu, Chung-Hsien, Jui-Feng Yeh, and Ming-Jun Chen. “Domain-Specific FAQ Retrieval Using Independent Aspects.” *ACM Transactions on Asian Language Information Processing*, Vol. 4, No. 1, pp. 1-17, March 2005.

Xu, Jinxi, Ralph Weischedel, and Ana Licuanan. "Evaluation of an Extraction-Based Approach to Answering Definitional Questions." In Proceedings of *the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2004),* pp.418-424, 2004.

Zhang, Zhuo, Lyne Da Sylva, Colin Davidson, Gonzalo Lizarralde, and Jian-Yun Nie. "Domain-Specific QA for the Construction Sector." In Workshop of *ACM SIGIR Conference*, July 29, 2004.

Appendix - Unified Medical Language System

The ontology which we used is Unified Medical Language System (UMLS). It is developed by NLM. The system integrates three knowledge bases: the Metathesaurus, the Semantic Network, and the SPECIALIST lexicon. We use the Metathesaurus to propose our method for understanding the meaning of the medical knowledge. The Metathesaurus preserves the names, meaning, hierarchical contexts, attributes, and relationships in the context form. We translate the Metathesaurus into the form of database because the database is good for searching. There is an instance in Table A.

Table A. Example of the hierarchical ontology

Concept	**Terms**	**Strings**
C0004238 Atrial Fibrillation Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**L0004238** Atrial Fibrillation Atrial Fibrillations	**S0016668** Atrial Fibrillation
		S0016669 Atrial Fibrillations
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation
		S0016900 Auricular Fibrillations