Radboud Universiteit Nijmegen


# Dynamic QA Systems Using Knowledgebase

## Using an Improved Accuracy Method Based on HowNet

Shiqing Hu

Nijmegen, August, 2013

Radboud Universiteit Nijmegen


# Dynamic QA Systems Using Knowledgebase

## Using an Improved Accuracy Method Based on HowNet

Author: Shiqing Hu

Student Number: s4200411

Supervisor: Prof.dr.ir.Theo van der Weide

Second Reader: Prof.dr.ir. Eduard Hoenkamp

*A thesis submitted in fulfillment of the requirements*

*for the degree of Master of Science*

*in*

*Information Science*

Nijmegen, August 2013

# Acknowledgement

This thesis would not have been possible without the support of many people. First and foremost, I would like to thank my supervisor Prof. Theo van der Weide. His constant scientific guidance and support were essential to me throughout my research. He is also a personal coach who always encourages me to break my own limitation and try to explore new areas. I am more than thankful for everything he taught me.

I would like to thank all my teachers during my study life, especially in my master life of Radboud University Nijmegen. They have taught me the necessary skills to become an educated and professional person, let me be able to do research indecently.

I owe special thanks to my boyfriends, Dr. Yiwei Gong. He always accompanies with me and encourages me since the day we met. He also showed me more details of how to write a qualified master thesis in English. The master thesis cannot succeed without his effort.

I also need to thank my dear friends and classmates. Studying with them I felt at home right from the first day of my master study. They all inspired me and helped advance my knowledge.

Last but not least, I would like to manifest my deep gratitude to my family for their endless support and sacrificial love. Thanks also go to many friends more than I could possibly mention here. Thank you all for reminding me that there is a wonderful life beyond the research.

# Abstract:

Nowadays, with the quick development of Internet, the amount of available information still is growing exponentially. This way the Internet has become an important resource to answer user's questions. The traditional search engine can help people search information conveniently to some extent, but it also may return so many relevant results mixed with a large number of irrelevant results. In this way, users can hardly get their desirable answers immediately. Consequently, people are in need of a new more intelligent information retrieval system which can satisfy their needs. Question-Answering (QA for short) system is such a new way of information retrieval. QA system provides us a with human-machine interface that support natural language. For one thing this mode conforms to the way how people ask questions, for another it has clear advantages over the traditional search engine on understanding the intent of a question. In addition, QA system in several functions more accurately than search engine because it returns the more accuracy answer.

This thesis presents an on-going research of the architecture and operation of a dynamic Question Answering system. The main results of this thesis are the architecture to handle the smart answering. This system is designed for Natural Language Processing in open domain based on Hownet knowledgebase, based on the knowledgebase, Natural Language Processing can get better effect; the system is available for learning ability. The ability is realized by the function of database updating. This system aims at high speed and accuracy on searching answers.

Three components were improved by improving their algorithm. In texts clustering, we proposed a new ant-tree algorithm to replace the traditional K-means algorithm. It speeds up the procedure of clustering. During the key words extension, we combine the classical algorithm with HowNet knowledgebase in order to keep the high accuracy of understanding natural language. In the answer extraction, we improved the obsolete algorithm by considering necessary factors.

Through testing a prototype of dynamic QA system, the improved algorithm can be proofed to be effective. In addition, the new architecture combined with the improved algorithms is the main contribution of this thesis.

**Key Words:** QA system; HowNet knowledgebase; Clustering; Learning ability; Improved algorithm.

# Table of Contents

# 1 Introduction

## 1.1. Question-answering System

Question-answering system is an application system to handle and research natural language (Waltz, 1978). It includes some basic skills of handling natural language and retrieving information, such as, lexical analysis, syntactic analysis, semantic analysis, and even index, retrieve and sort for large documents, etc. Question-answering system includes three parts, i.e., query, information retrieval and answer extraction.

Question-answering system is a typical application of information retrieval technology. It is often used to retrieve from text collections, and can answer open-ended natural language questions. Traditional question-answering system is often based on a fixed documents database rather than a complete internet-based database(Waltz, 1978) , therefore, it can only be applied in a certain domain. For example, as for the LUNAR of 1970s, geologists can look for useful information about moon taking advantage of it.

## 1.2. Question-answering System with the Interface of Natural Language and Foundation of Database

BASEBALL (Green in 1961) is the most famous question-answering system in early times which was designed for American football playoffs. For example, to answer users' questions like "Who did the Red Soxlose to on July 57"or "On how many days in July did eight teams play?". At this time, BASEBALL will analyze these questions using linguistic knowledge and then retrieve relevant information from database. Of course, there is lots of information about American football games in it.

BASEBALL can still be regarded as a relatively complex question-answering system considering from contemporary standard, because it analyzes questions from syntactic and semantic aspects. However, it also has drawbacks, i.e., firstly, it is based on a limited domain - baseball, secondly, it is based on structural database rather than text corpus. From this perspective, BASEBALL is the first good example among a series of question-answering systems with the interface of natural language and foundation of database. Such kind of system saves large amount of data in structural database and these data are available for users.

From the perspective of interface, the system uses question comprehensible interfaces which enable users to use their natural language to raise questions rather learn to use special language to retrieve database or know database structure.

LUNAR system (1973 Woods), an early question-answering system, is also designed according to the same principle. LUNAR system is designed to provide data for geologists researching moon to know, compare and estimate the chemical data about lunar rock and soil and its valuable data was achieved from Apollo Program. LUNAR system can answer 90% of geologists' questions about a certain domain, thus it is still a limited question-answering system.

Considering modern question-answering system, its limitation is that it is a structural knowledge database based on limited domain rather than an open unstructured text database. What should be cherished are the syntactic and semantic analysis toward questions, as well as the interaction between users and systems.

## 1.3. Interactive Dialogue System

Another famous question-answering system is human-machine dialogue system. In 1950, AlanTuring put forward an idea to treat Q&A as the test toward machine intelligence. He assumed that let a quizzer to ask questions toward an unknown entity (people or machine), then give possible answers according to entity's feedback.

Early dialogue systems, such as, SHRDLU (Wmograd in 1972) and GUS (Bobrow in 1977), are mainly designed to help researchers know relevant problems about human dialogue modeling. Although these systems use structural data as knowledge database, in fact, text corpus can also be applied in these systems. For these systems, their main problem ahead is taking advantage of the drawbacks generating from human dialogue, that is, pronoun repetition. Because machine cannot recognize what the pronoun really stand for, and it cannot figure out the ellipsis problem in oral English.

## 1.4. Question-answering System Based on Internet

From late 1980s, text processing changes from "rational form" represented by Chomsky to the statistical analysis of real text data and empirical knowledge induction. Undeniably, the change has something to do with the increasingly improved computer processing ability and accumulated text data. Respecting text language reality has become basic position and standpoint of text-based processing technology.

The successful development of search engine helps people get information more conveniently, and the internet information also increases in an unprecedented speed. A new problem comes out; people find that it is more and more difficult to find needed information accurately and rapidly through search engine.

Nowadays, Google and Yahoo are developed search engine. However, their search mode is key words form, and the return of search results are webpage collection concerned with

searching key words. Google and Yahoo rearrange the webpage according to some methods, such as, PageRank, but the rapid increasing information make them have two drawbacks:

Firstly, users spend large amount of time on reading the search results generated from search engine, and they also need to do further research to get needed results. Secondly, to get relevant information users need to change the question into a group of key words and arrange these key words logically (such as, using 'and', 'or'). In this case, users must be familiar with the logical presentation of key words. However, it is almost impossible for most common users. Consequently, inaccurate key words arrangement will generate inaccurate search results.

According to a research of MORI, only 18% users can get accurate information through search engine, while another 68% cannot. About 2/3 users are not satisfied with current search engine, hence to find a more efficient way to get information is a hot spot of current research

Google search engine make it possible to use internet documents. In 2002, Google has indexed 2.5 billion documents, and the project has not stopped yet. These search engines index information of different domains, what is more, these information updated continuously. Nowadays, a digital explosion has been taking place due to E-newsletter, E-journal, E-meeting Proceedings, website, digital communications, broadcasting, library and media. According to the research results of University of California, Berkeley, among the 2EB(Exabyte) non-repeating data generated a year globally, 93% of them are saved in digital form and which is 10 times the printed material throughout history, 25 times (about 8 Petabyte) the current Internet information data, 6000 times the data generated from earth observation system (EOS), 200,000 (about 10 Terabyte) times the printed collections in Library of Congress, United States.

Usually common users cannot give appropriate key words. That is why they prefer to use natural language to raise questions; and then the system analyzes their questions, searches for and returns needed information. Regarding to current search engines, they often treat questions of natural language as key words for query which return the unsatisfied searching results. At the same time, because users want the information returned to be accurate and concise, they often need to open those returned documents and skim over them to find information needed. It is very time consuming.

Supposing that there is a system which just returns a dialogue toward a natural language question, and the dialogue just contains answer. Or, the dialogue returned is exactly the right answer.

This is the open-domain question-answering system put forward in the late 1990s. Researchers of information retrieval domain promoted the question-answering system research and put forward a new way: get information needed from Internet and document collection using information retrieval, information extraction and natural language processing and treat the information achieved as data source so as to replace the knowledge base of

expert system.

With regard to questions about reality, users often just need a short answer. For example, the question-answering system receives a natural language question like 'Which river is the longest river of China?', in this case, an good answer should be "The Changjiang River" or "river Changjiang". If the traditional question-answering system can be expanded and search engine can be treated as data base, internet-based question-answering system will not only have natural language human-machine interface which is friendly and natural, but also give more accurate answers to users comparing with traditional search engine.

## 1.5. Thesis Structure

This thesis consists of eight chapters.

Chapter 1 presents the relevant background material and the brief history of QA systems. It also describes goal of research and the thesis structure.

Chapter 2 presents the literature study. It contains the understanding of question and answer. Besides, it also introduces the main technical methods and algorithm.

Chapter 3 discusses the system requirements of QA systems. The stakeholder analysis, environment human proportions and risks are included in this chapter. The requirements also predict different risks.

Chapter 4 presents the proposed system architecture. It is the main contribution of this thesis. It includes the framework of system and the introduction of HowNet knowledge database.

Chapter 5 introduces the natural language processing. It is the most important method in designing QA systems. It includes NLP procedure, shallow parsing and question analysis.

Chapter 6 describes the improved algorithm of QA systems. It involves cluster analysis, key words extension and answer extraction.

Chapter 7 discusses the evaluation methods of QA system.

Finally, the conclusion is presented in Chapter 8.

# 2 Literature Research

## 2.1 What is Question?

A question is an expression to request information.

While there are a number of different forms of question, this thesis primarily concerns factoid and definition questions as defined within the QA framework used within the Text Retrieval Conference (TREC).

Factoid questions are those for which the answer is a single fact. “When was Albert Einstein born?”, “How long is the Erasmus Bridge?”, and “Where is Google based?” are all examples of factoid questions. Questions other than factoid questions are not covered by this thesis. These include questions which can be answered by “yes” or “no”, such as “Is London’s population bigger than that of Paris?” as well as instruction based questions (e.g. “How do I make fish soup?”) and explanation questions (e.g. “Why did the Netherlands enter EU?”). Another type of question is list question which is closely related to factoid questions. List questions are factoid questions that require more than one answer. For example “What grapes are used in making wine?” is a list question. While list questions will not be covered in any detail in this thesis, most of the approaches to answer factoid question can also be used to answer list questions.

Factoid questions have attracted the main attention of recent QA research partly because they are the main focus of the QA evaluation which has been held annually since 1999 as part of TREC (Charoenpornsawat, Sornlertlamvanich, & Charoenporn, 2002). TREC QA evaluations (Ralph Grishman & Sundheim, 1996) suggest that the current state-of-the-art QA systems can answer at most 80% of factoid questions (although the median score in the same evaluation was much lower at only 17%).

Unlike factoid questions, definition questions require a more complex answer, usually constructed from multiple source documents. The answer should be a short paragraph which succinctly defines the definition which the user wishes to know more about. Good answers to definition questions should probably be very similar with an entry in an encyclopedia. For example, if the question asks about a person then the user will probably interest in the important dates in his life (birth and death), his major achievements in career and any other interesting stories of him. For an organization the definition should probably include information about its industry, time of foundation and founders, the size of the organization etc.

Definition questions have also been included in TREC QA evaluations. TREC QA evaluations (Ralph Grishman & Sundheim, 1996) suggest that the state-of-the-art QA systems can achieve an F-measure score of approximately 0.46, with the median score in the same evaluation

being approximately 0.18.

## 2.2 What is Answer?

There have been many attempts to define what constitutes a correct answer produced by a QA system. Several definitions of answer can be found in the TREC QA evaluations including:

• In TREC-8 (Charoenpornsawat et al., 2002) an answer was defined as a string of up to 50 or 250 characters in length which contained a correct answer in the context provided by the document;

• For TREC 2003 (Strassel, Przybocki, Peterson, Song, & Maeda, 2008) a response was judged correct if it "...consists of exactly a right answer and that answer is supported by the document returned."

• A number of studies have used the TREC questions and have defined an answer to be a text snippet which matches an answer judged to have been correct in the original evaluation and which comes from a document also judged to have been relevant (Roberts & Gaizauskas, 2004).

A precise and comprehensive definition of what constitutes an answer is difficult to arrive. Whilst an answer has to be correct to be of any use, this still leaves a lot of scope for different systems to present the same answer in different ways.

## 2.3 Technical Methods

### 2.3.1 Natural Language Processing

Natural Language Processing (NLP) is a domain of research and application that explores how computers can be used to understand and manipulate natural language text or speech (Chowdhury, 2003). NLP is an interdisciplinary research that might involve Computer Science, Information Science, Mathematics, Artificial Intelligence, Psychology, and so on. The input of NLP can be spoken language, text or keyboard input. In this thesis, we focus on the design of QA system and will not address the different ways of input for NLP. Simply, we assume that questions for QA systems are expressed in natural language and generated from text or keyboard input. Nevertheless, input from spoken language can be processed by adding by adding an application for speech recognition.

### 2.3.2 Shallow Parsing

In natural language processing shallow parsing has received much attention. Shallow parsing refers to a range of techniques that produce useful yet not complete syntactic analysis of text (Abney, 1997). The main goal of a shallow parser is to divide a text into segments which correspond to certain syntactic units. Although the detailed information from a full parse is

lost, shallow parsing can be done on non-restricted texts in an efficient and reliable way (Molina & Pla, 2002). In addition, partial syntactical information can help in solving many natural language processing tasks, such as information extraction, text summarization, machine translation and spoken language understanding.

Shallow parsing involves several different tasks, including text chunking, noun phrase chunking or clause identification. Text chunking consists of dividing an input text into non-overlapping segments. These segments are non-recursive, namely they cannot include other segments and are usually called chunks as defined by Abney. Noun phrase chunking (NP chunking) is a part of the text chunking task, which consists of detecting only noun phrase chunks. The aim of the clause identification is to detect the start and the end boundaries of each clause (sequence of words that contains a subject and a predicate) in a sentence.

Chunks and clause information in a sentence can also be represented by means of tags. Tjong Kim Sang et al., summarized several equivalent chunk tag sets for representing chunking. The IOB2 set Ratnaparkhi, uses three kinds of tags: B-X for the first word of a chunk of type X; I-X for a non-initial word in an X chunk (Molina & Pla, 2002); O for a word outside of any chunk. For clause identification, each word can be tagged with the corresponding brackets if the word starts and/or ends a clause, or with a null tag if the word is not the start or the end of a clause.

### 2.3.3 Text Cluster Analysis

Cluster analysis divides data into groups (clusters) that are meaningful, useful or both. If meaningful groups are the goal, then the clusters should capture the natural structure of the data. Occasionally, however, cluster analysis is only a useful starting point for other purposes, such as data summarization. Either for understanding or utility, cluster analysis has long played an important role in a wide variety of fields including psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning and data mining.
For information retrieval, the World Wide Web consists of billions of Web pages, and the results of a query to a search engine can return thousands of pages (Tan, Steinbach, & Kumar, 2005). Clustering can be used to group these search results into a small number of clusters, and each of them captures a particular aspect of the query. For instance, a query of “song” might return Web pages grouped into categories such as composers, lyricists, singers/stars, and production houses. Each category (cluster) can be broken into subcategories (sub-clusters), producing a hierarchical structure that further assists a user’s exploration of the query results.

### 2.3.4 Information Retrieval

When a user enters a query into the system, an information retrieval process begins. Queries are formal statements of information needs, for example search strings in web search engines.

In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.
An object is an entity that is represented by information in a database. User queries are matched against the database information. Depending on the application the data objects may be, for example, text documents, images, audio, mind maps or videos (Lashkari, Mahdavi, & Ghomi, 2009). Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

# 3 Requirements of Dynamic QA Systems

## 3.1 Stakeholders

In the open domain dynamic QA system, we have pointed out three kinds of stakeholders: Builder, User and Tester. Builder and Tester also have different roles. We define the stakeholders with the roles as the following table.

Table 1 Stakeholder of QA system

| Stakeholder | Role | Task |
|---|---|---|
| Builder | Software engineer | Building the QA system:<br>Building the question database, answering database, and result matching database.<br>Maintaining and updating the database.<br>Designing the algorithm of the QA system. |
| | Human Expert | Answering questions and producing knowledge.<br>Providing the answers of the questions which users put into the question database.<br>(So in other words, Experts are the builders of answering database.) |
| User | Main User | Consuming knowledge and producing questions.<br>Receiving the answers that they asked from the answering database.<br>Putting questions into the questions database.<br>(In other words, Main User is also the builder of question database.) |
| Tester | Systme Tester | Testing the validation and the effectiveness of the system. |
| | Quality Assurance Manager | Checking the validation of information, (both questions and knowledge). Moving the invalid information from the database. |

The whole system can be conceptualized by the following figure.

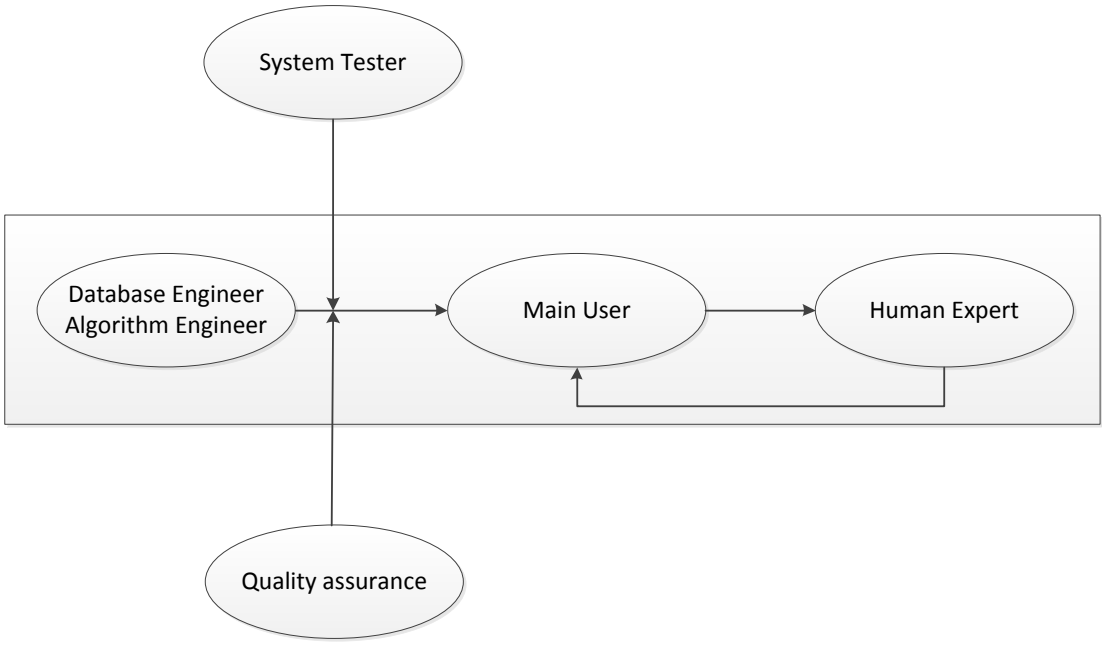


Figure 1 Role in the System

Software Engineer designs and creates the QA system. In the whole system development lifecycle, System Tester and Quality Assurance Manager keep on maintaining the system. When the system is launched, the Main User will judge the case (the provided answers of his/her questions). If Main Users are not satisfied with the case, Human Expert will provide assistance manually. In this process, the integrity and performance of the system will be improved.

The typical example of the dynamic QA system is wiki technology. 'Wiki' is not only a website, but also a technology and concept which is produced by internet. The key words used to describe wiki are 'open', 'collaboration', 'equality' and 'share'. Every stakeholder can participate in the process of the page building, and each one can produce questions and knowledge. At the same time, they can consume the knowledge and questions; they together develop the website from nothing to something, from rough to perfect, from isolation to mutual reference, and eventually the wiki becomes a high quality knowledge system.
Not only the building, but also the Quality Assurance helps in improving the system. When the knowledge or the questions in the database are out of the date, the Quality Assurance Manager moves invalid information from the database.

Actually, Quality Assurance Managers do not exist alone. In Wiki example, Users, both experts and receiver are can become Quality Assurance Managers. When a Human Expert sees some questions that are out of date, they can delete them from the database. For instance, in computer operation system fields, there were questions as "How much of a DOS system". In modern society, it is not valid question since DOS system is not used in most personal computers. The Human Expert can become the Quality Assurance Managers to delete this question. Of course, at the same time, they can become the QAs as the receiver to delete the answers. In this way, we can keep the size of database small, and keep it work normally in

case of break down by large data burden.

This is just in WIKI example, when we build our system; I prefer to separate the QA and users. In this management, we can keep the valid information all the time, control the data redundancy as much as we can and reduce the cost of maintenance as much as possible.

## 3.2 Environment

The first aspect of environment is web-based: The Web, which is home to billions of pages of electronic text, is orders of magnitude larger than the TREC QA document collection, which consists of fewer than 1 million documents. This is a resource that can be usefully exploited for answering question.

QA system is a web based application that runs in a cloud service. The source of the system comes from web and the system is used at web. The system can realize the semantic retrieval of information source based on web. We take advantage of the tremendous data that the Web provides as the backbone of the smart QA system, from creating, updating and testing process.
The second also the most important environment aspect of our system is that the Main User and Quality Assurance Manager can access the system from anywhere at any time.

Alternatively the system can be implemented in client-server model. In this case, we have to build and release a client application that is convenient for all the stakeholders.

The structure can lead user access to the database from the client very easily. It is especially important for tester that they can keep the system normally anywhere at any time. Of course it is also convenient for users who want to get knowledge.

## 3.3 Human Proportions

For the QA system, the end product will be a web-based application. The user interface is simple so that the end user can easily find out the functions he/she wants. Because of the simple interface design, the end user will feel comfortable when using the system.

There is cooperation in the whole procedure, which builders, experts and users collaborate together. It means that, experts and users are also the builders at once. And the experts can be the users at the same time the users are also the experts.

What the mean idea of building this smart QA system is realize the target of “knowledge sharing”, with not only simple answers but also experience files. In that way, we hope one person can have different roles, for example, experts and receivers can become each other. In this definition, the knowledge can be produce, consume and shared maximally.

Besides, for human brain thought for question, the principles are from easy to hard, from widely to professional. So considering this fact, we set the questions in different levels, that not only easy for builders to build the system, but also easy for experts to produce the knowledge.
Here I would like to show an example.

There is a utility company, which delivers electricity at private home. The field workers have to do: add new customer and repair the electricity when there are some problems.
We have the internet application that services for the field workers only.
Then we have the question levels such as:
(1) Please introduce some device.
(2) What's the procedure of reparation?
(3) What's the environment information of users when using the specific device?

In this case, we can notice that for question 1, it is simple question. For question 2, it is more complex than question 1, and then for question 3, about the specific information, it is totally complex for workers to answer.

## 3.4 Risks and Cost

Every project in processing will have some risks, and for every risk, we will take some actions to remedy the bad influence. Of course, before we start project, we will evaluate probability for every risk, and compute the cost of action to resolve the consequences of risks.

Table 2 Risk and Cost

| *Risk* | *Probability* | *Action* | *Cost of action* |
|---|---|---|---|
| *The builders cannot get enough knowledge. from industry* | *Low* | *Switch the specific industry to the similar ones* | *Low* |
| *There is large quantity of knowledge. which the database cannot deal with* | *High* | *Using pre-processing to choose more useful knowledge source* | *High* |
| *There will be a mistake in estimating the working building time.* | *Very High* | *Adjust the time management during the processing* | *Very high* |
| *The project is too expensive to* | *Very High* | *Using the cheaper* | *High* |

| | | | |
|---|---|---|---|
| *roll out* | | *development tool instead of expensive ones* | |
| *Server breaks down* | *Low* | *Asking experts for help* | *0* |
| *People can make mistakes* | *High* | *Change the workflow or behavior to correct the mistake.* | *Dependence.* |

For time management and financial management, we evaluate them by the Pareto principles. The Pareto principle (also known as the 80–20 rule) states that, for many events, roughly 80% of the effects come from 20% of the causes.

It is a common rule of thumb in business; for example "80% of your sales come from 20% of your clients". Mathematically, where something is shared among a sufficiently large set of participants, there must be a number k between 50 and 100 such that "k% is taken by (100 − k)% of the participants". The number k may vary from 50 (in the case of equal distribution, i.e. 100% of the population have equal shares) to nearly 100 (when a tiny number of participants account for almost all of the resource). There is nothing special about the number 80% mathematically, but many real systems have k somewhere around this region of intermediate imbalance in distribution.

In the software condition, from this principle, we can predicate that the risks of time and financial are very high.

For evaluation of servers break down, we calculate in this method:
Let p as the probability of server breaks down.
Let 1-p as the probability of server functioning.
Then if the system consists of 2 servers:
The probability of both servers functioning is: (1-p)(1-p)
The probability of servers completely break down is: 1-(1-p)(1-p)
If there are 3 servers:
The probability of both servers functioning is: $(1-p)^3$
The probability of servers completely break down is: $1-(1-p)^3$
Of course, we can calculate the probability of situation of more servers.
From this computing, we can know that more servers, less probability of break down.
So in dynamic QA system, we will use at least two servers to insurance cost in reparation, because the cost of reparation is higher than building.

# 4. The Proposed Architecture

## 4.1. Problem Description

At the initial period of network development, the appearance of search engine did provide a sound platform for online searching, which indeed facilitates users. However, with the rapid development of network and information technology, the explosive growth of online information brought some problems to the usage of search engine. Current search engines usually adopt the method of keyword search, which only enables several functions including recording large amounts of websites for searching, matching website record through keywords, or sequencing search in accordance with keywords' matching degree. When a user enters a question directly in the search engine, many files with the content of questions might be listed, but not always with answers. However, with the help of QA systems, based on a search engine, more intelligent operations can be carried out, for example, understanding user requirements by analyzing questions, organizing new search to imitate the context of answers, or extracting answers from retrieved files.

As traditional search engine relies on keyword searching, using traditional search engine in an accurate way is a complicated operation which requires users have relatively higher levels of technique to precisely describe keywords. In addition, the key issue of searching is the pickup of keywords and organizing them properly. However, people want to get more accurate information in a convenient way within a short time. Therefore, traditional search engine cannot satisfy this requirement. To satisfy this requirement, the QA system proposed in this thesis provides a more accurate and rapid method to obtain information. The design concept, function mechanism and expectance are all different from current keyword searching. It is information retrieval at a higher level. Nevertheless QA systems will not replace the traditional online search engines. They are a new generation of search engines, which can understand natural language and facilitate users to obtain information during a more people-friendly interaction procedure. This will revolutionarily change the way how people used to obtain information from computers and internet, with great theoretical and practical value.

QA systems can process a question in natural language in sentence, and directly feedback required answer to user, instead of the hyperlinks of relevant websites. In this sense, QA systems are more convenient, and quickly return the answers asked by users. That is to say, QA system is the new generation of search engine. As for QA system, users don't need to disassemble their questions into keywords. Instead, they just enter the whole questions to the QA system. Combined with the natural language processing techniques, QA system can provide users answers through the understanding of questions. QA systems like an expert with all kinds of knowledge in different fields, who can answer any question rapidly and accurately. For example, if the question put forward by user is "What is the Bluetooth technology?", then QA system will directly pop up answers as "Bluetooth technology is the short distance

wireless connecting technology which takes place of cables or wires on portable or fixed electronic devices". It is obvious that QA systems are more convenient, faster and more efficient than traditional search engines.

The QA system proposed in this thesis is different from an expert system, information retrieving system or information extraction system in essence. It is not the simple transformation of traditional technology, but a researching topic with high theoretical and practical value.

## 4.2. Framework

### 4.2.1. Traditional QA system

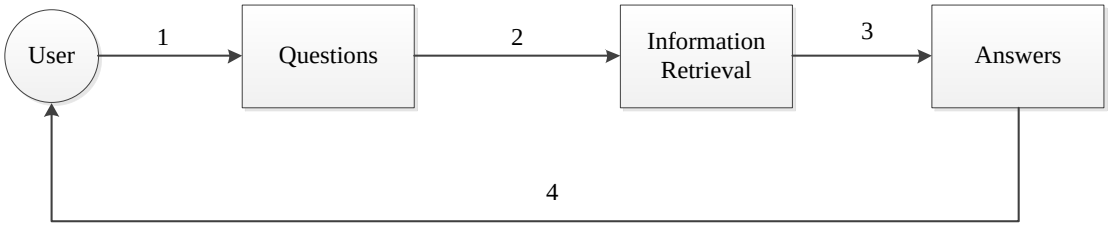


Figure 2 Traditional QA System

The traditional QA system usually contains the process likes Figure 2.
System starts with users. Users ask a question, and put it into the system. Then we have the process:
1: Deal with the original questions, usually using natural language processing, and then we can get the middle language of the questions. Questions after the dealing can be stored in the question database.
2: Using text clustering to deal with the question in database. Through this way, we can find the feature of texts. After that, we can put the irregular questions into different group by feature. In this way, we can obviously speed up.
3: After dealing with questions, we now can do the information retrieval, this stages contains both key words extension from question features and answer extraction from the knowledge base.
4: Finally, the system gets the answer which fit for the question, then, returns it to the end users.

## 4.2.2. Dynamic QA system

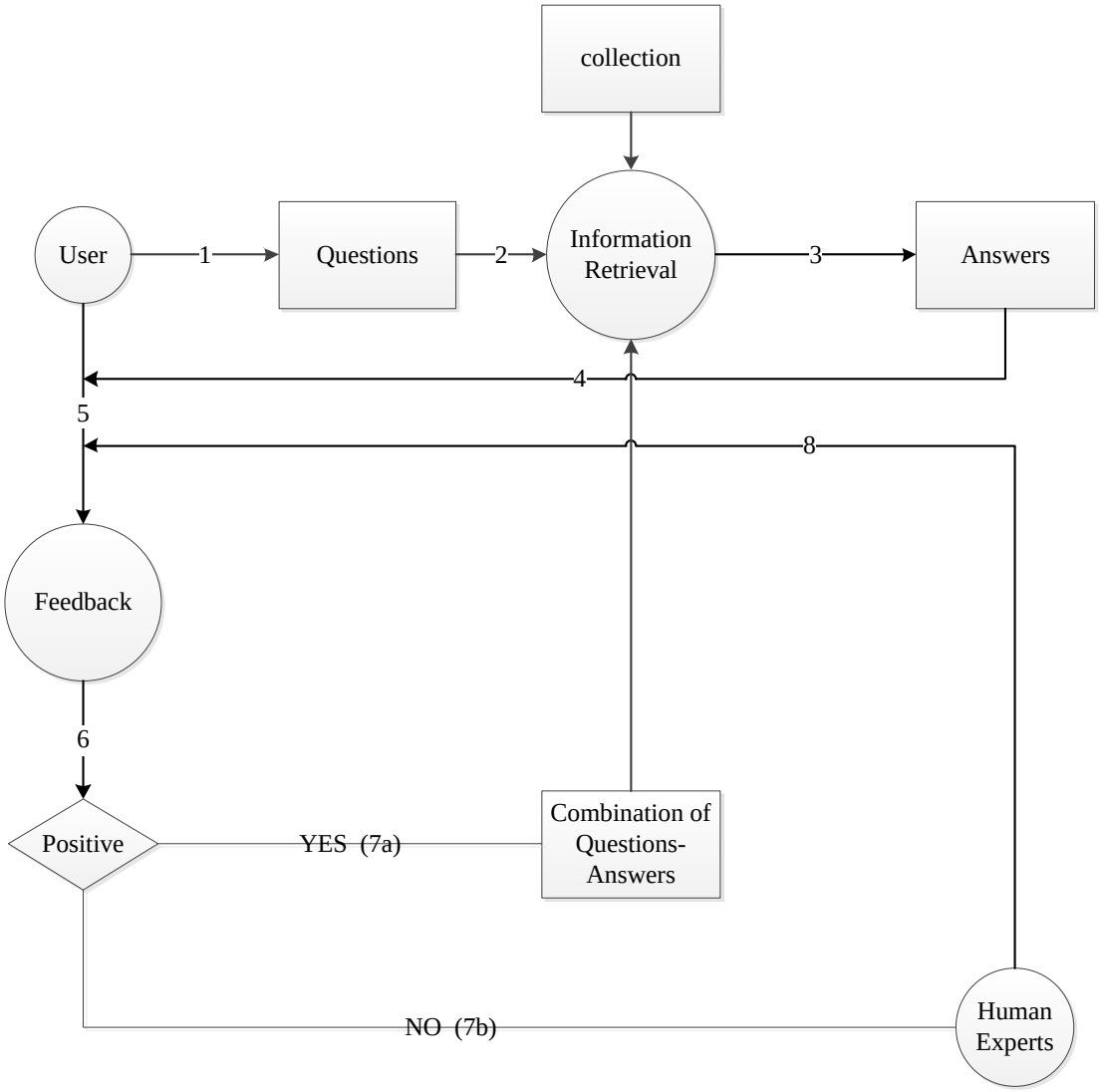


Figure 3 Dynamic QA System

We make some development base on the traditional one. Now see the Figure 3, the dynamic system adds two steps compared with the previous system:

5: After the answers return to end users. We should make the judgment whether users are satisfied with the answers. If they say yes, then we store the combination of question and answer into the matching result database. If they say no, then we send the question to do the artificial answering, that we call it user 2.

Actually, there are two groups of users in our system. They deal with the different information. Also there are two groups of information in our system. One group is question, another is answer.

User1 put the question into the system and get the question from the system; User2 put the answer into the system and get the question from the system.

6: After the artificial answering, we send the answer back again to user1. Let user1 say yes or

not. Still, if yes, store the result, if not, we do the cycle again of step 5.
In order to get the better performance, we make more development and finally get the new framework of QA system. It aims at high speed, accuracy and the learning ability.
We can see from Figure 3, the bold face arrow shows the new path of finding the answers. After first step, we go directly to the matching results database to find the combination of question and answer. If we can find it, return it to user, if not, go straight to information retrieval. In this way, we can obviously speed up.

### 4.2.3. Main Algorithm

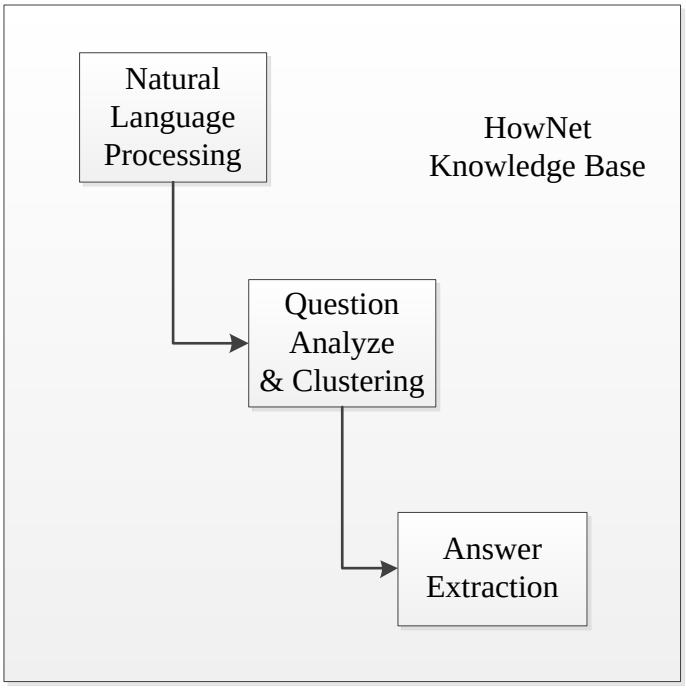


Figure 4 Main Algorithm

In Figure 4, we can find that the main algorithms of building system are together based on HowNet Knowledge Base.

First we do the Natural language processing, after that we do the question analyze and clustering, finally we do the answer extraction and information retrieval.

## 4.3. Hownet Knowledge base

### 4.3.1. What is HowNet knowledge base

HowNet is an on-line common-sense knowledgebase unveiling inter-conceptual relationships and inter-attribute relationships of concepts as connoting in lexicons. To put it simply, relationship is the soul of HowNet, as well as the world knowledge. The relationships that represent knowledge can be divided into two categories: Concept Relationship (CR) and

Attribute Relationship (AR) (Dong, Dong, & Hao, 2010).

### 4.3.2. How to build the HowNet knowledge base in the QA system

Hownet is a knowledge base that explains the relationship between concepts and the relationship between the properties of concepts in both Chinese and English. Hownet places the concepts of all vocabulary in the objective world under four categories, namely, entity, event, property, and property value, and deciphers concepts through original sememes. Original sememe refers to undividable semantic units.

Hownet is home to about 2,200 original sememes. Original sememe contains eight pairs of relationship, namely, hyponymy, synonymy, antonymic relationship, appositive relationship, property-host relationship, component-wholeness relationship, material-finished product relationship, and event-role relationship. Of all these original sememes, hyponymy is the core. Hyponymy is placed in a corresponding semantic document in the form of tree structure.
What is original sememe? It is difficult to define it just as it is difficult to define what word is. However, this does not prevent people from using it. Generally speaking, original sememe is the most basic and the smallest unit that can no longer be divided. For example, although "man" as a very complex concept contains a number of properties, it can be considered as an original sememe. We can envisage that all concepts can be divided into all kinds of original sememes. Meanwhile, there should be a limited original sememe set in which original sememes combine to form an unlimited collection of concepts. We can control and use the limited original sememe set to describe the relationship between original sememes and the relationship between properties. In this way, we can establish the knowledge system we envisage.

Hownet adopts Knowledge Database Mark-up Language (KDML) to describe the definition of a concept. KDML uses the semantic mode DEF to define a concept. DEF describes in detail the semantic properties of words and expressions. For example:

```
Birstday: DEF={time: TimeSect= {day} ,  {ComeToWorld:time={~}}}
```

In Hownet, a complete concept must be enclosed by a pair of braces ({}). The left brace means the beginning of the description of a concept while the right brace means the completion of the description. The left and the right braces must appear in pairs, in other words, they must match each other from the beginning to the end. The first original sememe of the DEF is the main original sememe of the concept, indicating the category to which the concept belongs. What is behind the colon is the explanation of the previous original sememe. This kind of explanation reflects seven kinds of relationship apart from hyponymy. Like the example above, there are two specific modes, namely, first dynamic role name={the description of a certain concept}, and secondly dynamic role= {~} is added to the concept of

event to indicate that the concept plays the semantic role of the event. In Hownet, a total of 90 dynamic roles or semantic roles are defined to express these seven relationships.

Hownet places all the events in the objective world under 817 most basic categories, that is, 817 event original sememes, while the number of property original sememe reaches 247, that is, 247 types of generalized properties in the objective world. Meanwhile, in Hownet, almost every event original sememe is equipped with a semantic role framework to explain the matching capability of its own. The necessary roles of that original sememe are listed in the framework. These roles are indispensable to each other, otherwise, that event cannot be validated. In other words, when a certain event occurs, all the roles in the framework must be involved. For example, the semantic role framework of "engage|从事" is {agent={*},content={*}}, that is, when such type of event of "engage|从事" occurs, there must be implementers and contents.

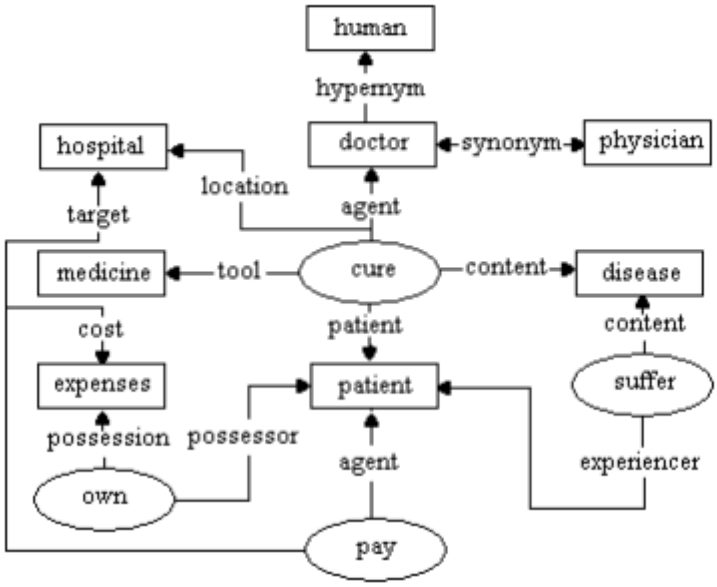


Figure 5 Concept Relation Net (CRN) of "doctor"(Dong et al., 2010)

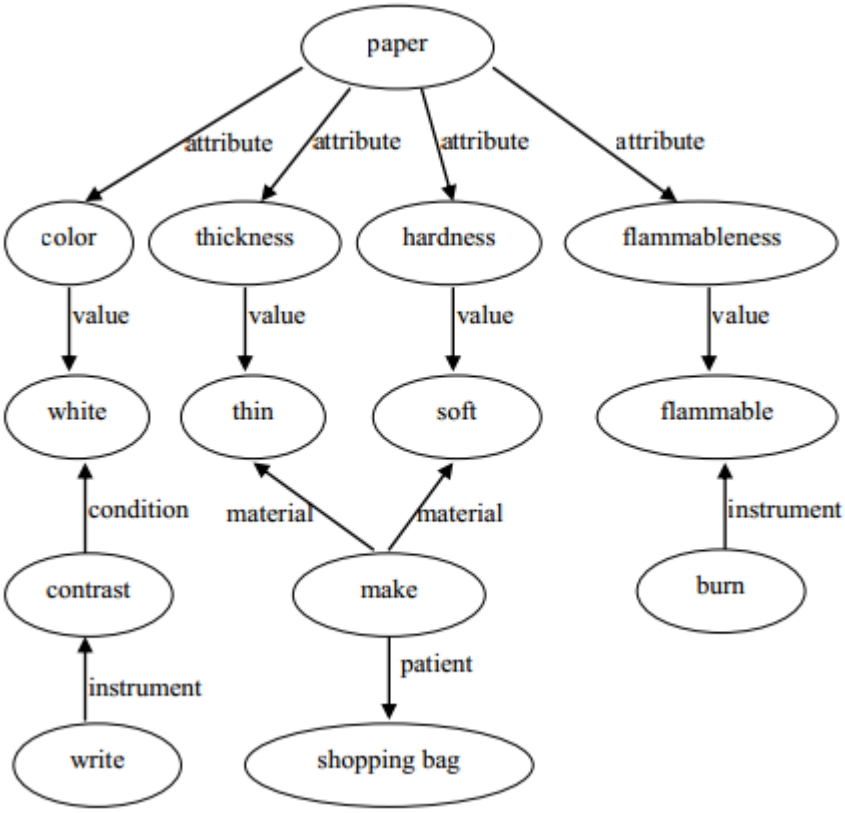


Figure 6 Attribute Relation Net (ARN) of "paper"(Dong et al., 2010)

# 5. Natural Language Processing

## 5.1 What is natural language processing

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. NLP is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications (Liddy, 2009).

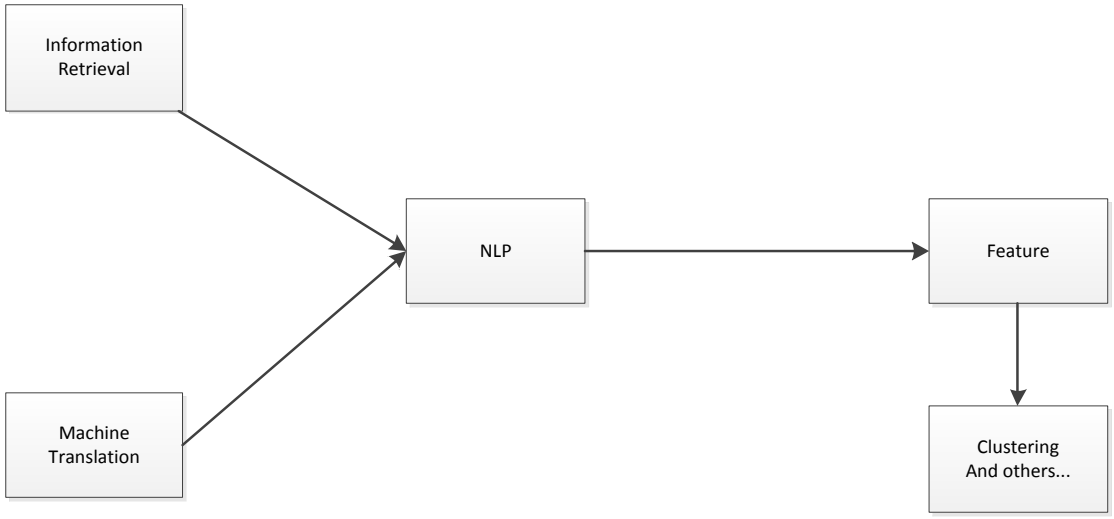


Figure 7 Overview of Natural Language Processing

If the QA system accepts questions described in natural language instead of computer language, the system is user-friendly. Such a QA system has to adopt Natural Language Processing (NLP) technology to come up with smart user-computer interaction. Most databases' natural language interface processing can be divided into the following four stages:

1. Natural language understanding
   What is the meaning of the question? This is the question which the first stage is to answer. In other words, grammatical, semantic and pragmatic analyses are made to transfer a user's natural language problem into a standard, definite (ambiguity-free) intermediate form. Although different systems have different intermediate forms, most of them are in the form of predicate logic.
2. Query generation
   How can I answer it? This is the question that the second stage is to answer. It transfers the definite intermediate forms into a sentence that has the formal query language required by the data base system, such as Structured Query Language (SQL) or relation calculus language. Or, this stage transfers the definite intermediate forms into a group of routine calls that execute queries.
3. Query execution
   What's the answer? This is the question that third stage is to answer. It does not belong to the interface category.
4. Result processing
   How can I present the answer? This is the question the fourth stage is to answer. First, an appropriate way, which could be a graph, an icon or a list, must be found to present the answer. Secondly, other useful information should be provided to make the system more natural and friendly.

From the four stages mentioned above, we find that the first stage is the core of natural language interface system. The mission of the first stage is to undertake grammatical and semantic analyses. Language is a tool to describe human being's thinking. In a sense, the

understanding of language equals the understanding of human thinking. So, language understanding is very difficult.

Natural language processing in information retrieval and machine translation is the mainstream in the current research.

### 5.1.1 Machine translation

The conventional method of machine translation is Rule-Based Machine Translation（RBMT） (Charoenpornsawat et al., 2002). RBMT mainly relies on a data base of language translation rules generalized by linguists. First, the analyses of grammar, semantics, morphology and part of speech are made. Then, judgment and deduction are made in accordance with corresponding grammatical rules. Finally, equivalent sentences in the target language are generated, that is, the translation results. RBMT is made up of several phases, including looking up dictionaries, simple analysis of the source language, the generation of the target language, and the target language forms and sequence adjustment. Throughout the process, no distinct hierarchy and module can be found.

In terms of translation patterns, traditional machine translation methods include direct translation, inter-lingual approach, and transfer approach.

(1) Direct translation
Direct translation refers to the method of directly translating the words or phrases of the source language into the corresponding words or phrases of the target language with the support of dictionaries. However, it may lead to imprecise translation results due to the neglect of the grammatical structures and semantic rules of a sentence. It is the most primitive translation method. Since the 1970s, as machine translation technology has been advancing, researchers have realized that the differences between a source language and the target language are not confined to just vocabulary, instead, the differences have extended to syntactic structures. Syntactic analysis technology must be introduced to obtain more comprehensible translations.

(2) Inter-lingual approach
The source language is analyzed and then transferred into a general intermediate language, from which the target language is generated. This is called inter-lingual approach. Theoretically speaking, it seems to be a very convenient method used in a multi-linguistics machine translation system. However, in practice, it will be very difficult to establish a general intermediate language that is independent of natural languages on one hand and is able to express these natural languages on the other hand.

(3) Transfer approach
Transfer approach adopts two types of internal expression forms and comprises three stages. At the first stage, the source language is transferred into its own internal expression. At the

second stage, the internal expression of the source language is transferred into the target language's internal expression. At the third stage, the target language is generated in line with the target language's internal expression. All the transfers at the three stages are based on a rule data base. Transfer approach is the current mainstream rule-based translation method.

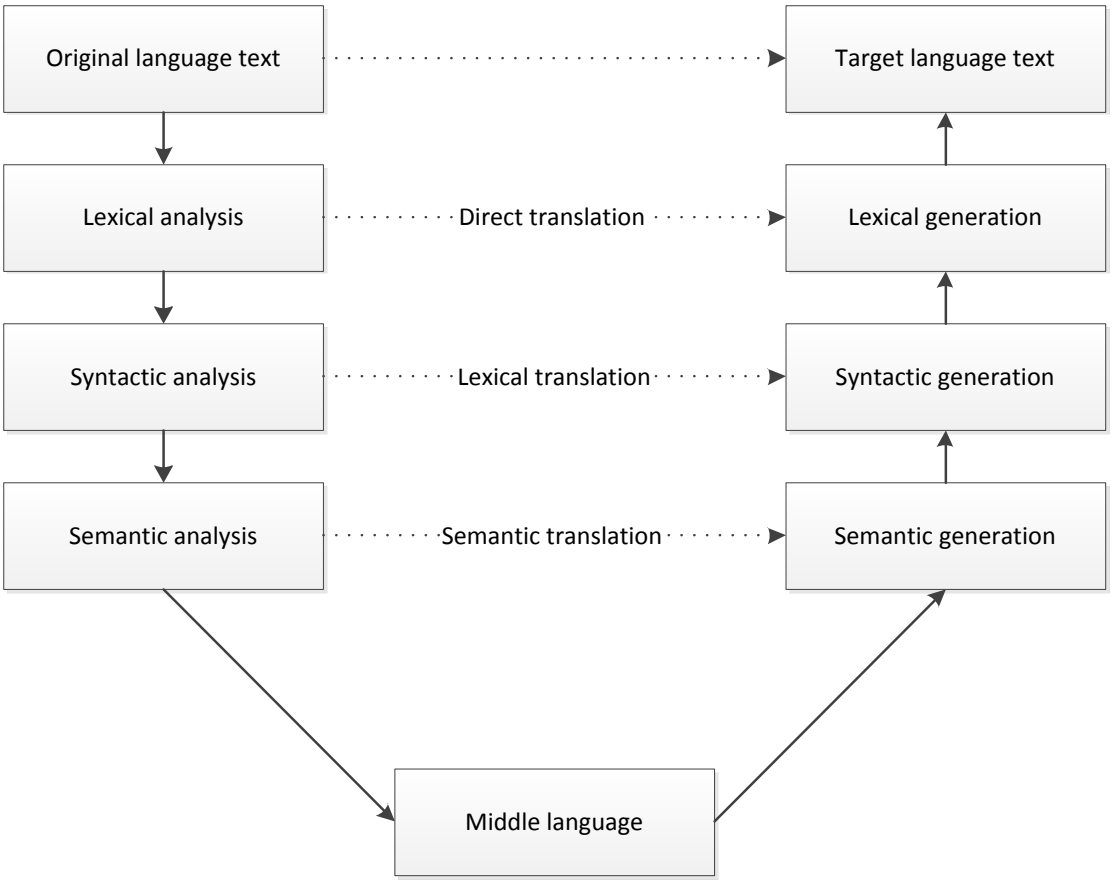


Figure 8 Overview of Machine Translation

## 5.1.2 Information Retrieval

Advanced natural language processing technology includes syntactic analysis, phrase recognition, named entity recognition, concept extraction, anaphora resolution and word sense disambiguation (WSD).

(1) Phrase recognition

The recognition of phrases in inquiries and documents relies on either syntactic analysis technology in natural language processing or statistical methods. Whether phrase recognition technology can be effectively applied to information retrieval mainly depends on pertinent recognition technology, phrase types and matching strategies. Recently, progress has been made in the application of phrase recognition technology.

Nie and Dufort integrated phrases, which serve as appended units, to traditional word-based indexes. They placed phrases and words on different vectors, and worked out the similarity between inquiries and documents, and then weighted them. The experiment results from

TREC6 and TREC7 data collection indicated that such method significantly improved the retrieval precision.

Nie and others considered various phrases in inquiries and documents, such as proper nouns, dictionary phrases, simple phrases and complex phrases. The phrase recognition technology they adopted is very flexible in that a phrase can be recognized once all the words making up the phrase appear in a window of a certain size. Every type of phrase has different corresponding size of window, which can be figured out through the approach of a decision tree. The intimacy of all the words making up a phrase should be computed. A phrase will be selected and used on the condition that its intimacy is above a certain threshold value.

(2) Named entity recognition

Named entity refers to a kind of special phrases that mark a certain concept or entity, such as proper nouns, names, geographic names, organization names, etc. Obviously, named entities contain more precise information than words and general phrases. However, the application of named entities does not result in more effective information retrieval. It is because on one hand named entity recognition technology is faulty in itself. On the other hand, researchers are confused about how to partially match named entities. For example, what weight should be given to Bill Clinton" and "Clinton"? It reminds us of the problem encountered when the precision of participles is very high. An approximation method can be adopted to solve this problem.

(3) Concept extraction

Concept is a kind of special phrases that are more common than named entities. Now that named entities mark a certain concept, they are also considered to under the category of concept. However, concepts contain more phrases beyond the scope of named entities, for example, information retrieval. Nevertheless, concept extraction does not effectively improve the result of information retrieval. Researchers are wondering whether it is necessary to use concepts in information retrieval, how to use concepts, and how to standardize different concepts that express the same meaning ( for example, 95%， 0.95 and percent 95).

(4) Anaphora resolution

Anaphora resolution technology helps pronouns in a document or phrases with no clear reference identify the things they refer to. For example, anaphora resolution technology can be used to specifically point out that "Mr. President" is used to refer to "Bill Clinton" and what "he" means in "He denied all responsibility". Since anaphora resolution technology can eliminate those indefinite modes, it seems to be conducive to information retrieval. However, it is not the truth. Anaphora resolution technology fails to significantly improve the effectiveness of information retrieval too. It is because on one hand anaphora resolution is faulty in many aspects in itself. And on the other hand, pronouns and phrases without definite reference do not seriously affect the results of information retrieval.

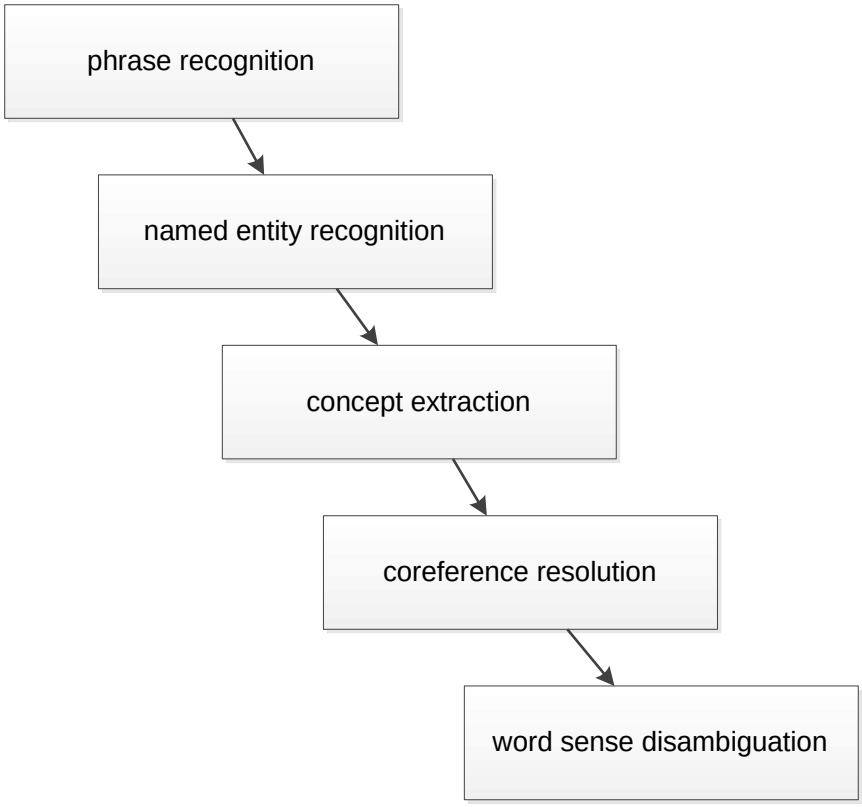


Figure 9 Procedure of NLP

## 5.2 OpenNLP

In this dynamic QA system, We choose the Toolkit of NLP to establish the progress of natural language processing.

The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.

It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. These tasks are usually required to build more advanced text processing services. OpenNLP also includes maximum entropy and perceptron based machine learning (Apache, 2010).

We chose OpenNLP Tools as an open-source tool suite which contains a variety of Java-based NLP components. Our focus is here on five "core" NLP components, viz. sentence detection, tokenization, POS tagging, chunking and parsing. OpenNLP is a homogeneous package based on a single machine learning approach, viz. maximum entropy (ME) (Buyko, Wermter, Poprat, & Hahn, 2006). The rationale behind ME for any collection of facts is to choose a model which is consistent with all the facts but otherwise as uniform as possible. Each OpenNLP tool requires an ME model that contains statistics about the component's default features combining diverse contextual information such as words around end-of-sentence boundaries

for the sentence splitter or word/tag combinations in five-word/tag-window for the chunker. The components are partly based on publications like the chunking model described by Sha & Pereira as well as the part-of-speech tagger and the parser described by Ratnaparkhi. For training and testing of all OpenNLP components we considered, the above-mentioned two major biomedically annotated corpora, viz. Genia and PennBioIE, were employed which currently contain the most elaborate annotations relevant for syntactic analysis in the bio field. We performed ten-fold cross validation for all OpenNLP tools on both corpora (Buyko et al., 2006).

## 5.3 Shallow parsing

Question-answering system is an application system of natural language processing research. It requires the support of Lexical, syntactic, semantic and pragmatic in natural language processing. Question-answering system performance will be better and better with the deepening of research, as the same time, study on question answering system will promote the development of natural language processing.

By the test results of TREC QA, it says that if we want to improve the performance of question answering system, relying on lexical and Syntactic Analysis is not enough, we must undertake an analysis of the Semantic and even Pragmatic aspects. In order to improve the treatment of question answering system, this chapter introduces the shallow parsing analysis in question answering system.

Shallow parsing is some kind of formal representation reflected the meaning of the sentence ( or semantics ) according to the syntactic structure of sentences and meaning of each word in sentence. For example the sentence: "Tom ate the apple " and " apple was eaten by Tom ", although they are different in form, but says are consolidated in the form of the idiom meaning to; "eat ( Tom, Apple ) "

Shallow parsing is a level of Natural Language Processing which is on top of Syntactic Analysis, the basis for computers understanding language, so it is very important.
Judging from the Application of Natural Language Processing, Whatever it is information retrieval, information retrieval, machine translation, automatic summary, or the human - machine interaction, first of all, it need understand the language and after you determine the correct meaning about the language, then follow up and get the results.
Judging from the development of Natural Language Processing, Syntactic Analysis can not achieve the satisfactory results in practical application, so researchers are turning to semantic study, proposing various theories of semantics.

The main objectives pursue of scholars engaged in researching on Natural Language Understanding is to make a correct shallow parsing of Sentences. However, after decades of development, there is not much use on learning method to get a detailed Semantic Understanding of knowledge. A feature of Natural Language is full of ambiguity. When the effect of parsing is less than satisfactory, shallow parsing of natural language uses semantic

knowledge, it will be benefit to solve the syntactic ambiguity problem which cannot be solved, in order to get a better understand of language. The function of shallow parsing of Question Answering System mainly show in:

(1) Word Sense Disambiguation
The same form of ambiguity word is very common. Even if one word has one part of speech, it can also have different senses, at this time only use syntactic knowledge to solve problems is Limited. Sense is in the area of the meaning of semantic, in the process of matching words with other words as well as in relevant context, words will meet certain constraints of the semantics, an important application of semantic knowledge is to resolute sense disambiguation problem of questions.

(2) Syntactic Disambiguation
Composition of Syntactic component is highly flexible, and it is lack of morphological change, which makes the analysis of Syntactic structure is very difficult and produces a lot of errors results, in addition, Syntactic Structure may not have the correct Logical Meaning by Syntactic Analysis. Lawfully semantic knowledge can be used to test the Syntactic Structure and exclude incorrect meaning of Syntactic Structure.

(3) Acquisition of Semantic Relations
Understanding questions not only need determine the meaning of the words in question, but also to determine the logical relationships between words in order to get the correct semantic information exactly. In the shallow parsing using semantic knowledge can help provide a semantic relationship between the components of the language fragments, and have a better understanding of the purposes of the users.

## 5.4 Question sorts

Briefly speaking, question classification means that in accordance with certain classification standard, it defines a category collection, and judges the category of questions into a certain collection based on some calculation. In the angle of math, question classification is a mapping process which maps questions without clarified category into existing categories. This mapping process can be one-to-one mapping, or one-to-several mapping, because questions are usually connected to many categories. This can be explained in the following math formula:

$$F: A \rightarrow B$$

A is the collection of questions waiting to be classified, B is the category collection in the classification system; the mapping rule of question classification: in accordance with the mastered data information of several sample of each category, the system should conclude rule of classification, establish distinguish formula and rules; then in face with new text type, the system can judge category of questions following the above concluded rules.

Question classification aims to make sure the semantic categories of answers and strategies adopted to analyze questions. Question classification is the most important and key step in the understanding process of questions, exerting great influence on further analysis. In most QA systems, major errors occurred are aroused by the incorrect question classification. Only making sure question type and answer type which need to be searched can answer put forward after applying some special strategies to analyze questions and search.

The classification method adopted in this paper is the method based on rules, putting forward the combination of interrogative and question focus to classify questions. In the first place, it defines following eight categories of questions.

Table 3 Question categories

| Question categories | Examples of interrogative | Expected answer | Examples of questions |
|---|---|---|---|
| About people | Who | Name | Who put forward evolutionism? |
| About location | Where | Location | Where host the 2008 Summer Olympic Games? |
| About time | When | Time | When is the first time that human landed on the moon? |
| About quantity | How much | Number | What's the area of Taiwan? |
| About definition | What | Explanation | What means artificial intelligence? |
| About method | How | Explanation | How to look energetic every day? |
| About reason | Why | Explanation | Why does bird flu happen? |
| Others | … | Complicated | What brands of famous clothes are there in the world? |

The first four categories represent inquiring people, location, time and quantity respectively, covering the most questions based on facts. These are key categories that QA system recognizes and processes as well. Correct answers of these questions are always found by named entity recognition and information extraction technology.

There are three more categories of questions inquiring definition, method or reason, with answers might be a paragraph or several paragraphs. Answers to these questions can't be provided through recognizing named entity, but more advanced technologies as paragraph understanding, automatically summarizing. The last category is others which are questions

can't be recognized as some certain categories of aimed questions. As for such questions, no concrete rules can be drawn out. Therefore, it will adopt probability classification which needs to collect large amount of questions as training corpus, count probability of each question through procedures, and then pick up question classification with largest probability as this category. Question classification module in this system will do the utmost to classify questions put forward into one of these eight categories. But in the following procedures, system will only focus on the first four categories.

The recognition of question classification mainly depends on interrogative in sentences. However, each interrogative has varied recognition ability to question categories. For example, if there is interrogative as "where" in sentence, it's easily to tell that such question category is which inquiring location; if there is interrogative as "who" in sentence, it's easily to tell that such question category is which inquiring people. For the convenience of narrating, such interrogative is collectively named as exclusive interrogative. However, if there is interrogative as "what", "which" in sentences, the question category can't be told just by such interrogative. Because many question categories can include such interrogative, which can be named as general interrogative. As for general interrogative, it still needs to combine question focus to make sure question categories.

# 6 Improved Algorithm

## 6.1 Cluster Analysis

### 6.1.1. Definition and Why It is Useful

Cluster analysis is a vital procedure in designing QA system, so there is a question, what's the meaning of Cluster Analysis?

There are a set of objects, we want to get a way that objects in the same group means cluster are more similar to each other than to those in other groups or clusters. This is the main assignment of exploratory data mining.

Cluster analysis someway can be an optimization problem. Small distances among the cluster members of groups, areas of the data space, intervals or particular statistical distributions, they are the notions of clusters. The algorithm and parameter settings depend on the individual data set and intended use of the results. Cluster analysis is an iterative process of knowledge discovery. The process involves trial and failure. It is always necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Now, the simple process of cluster analysis can be presented by the following diagrams.

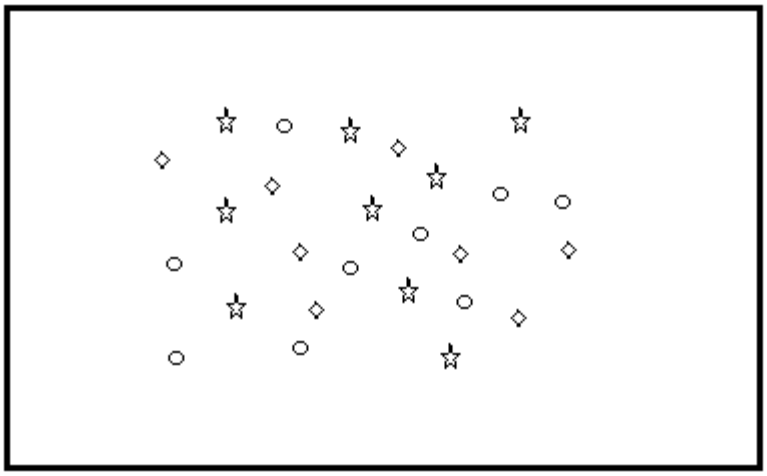

Figure 10 The initial point without clustering

Figure 10 shows that, there are some different kinds of points located in space. Each point has its own feature by shape. The intention of clustering is to do the classification. Such behavior can classify the points which have same shape together. The results can be shown by Figure 11.

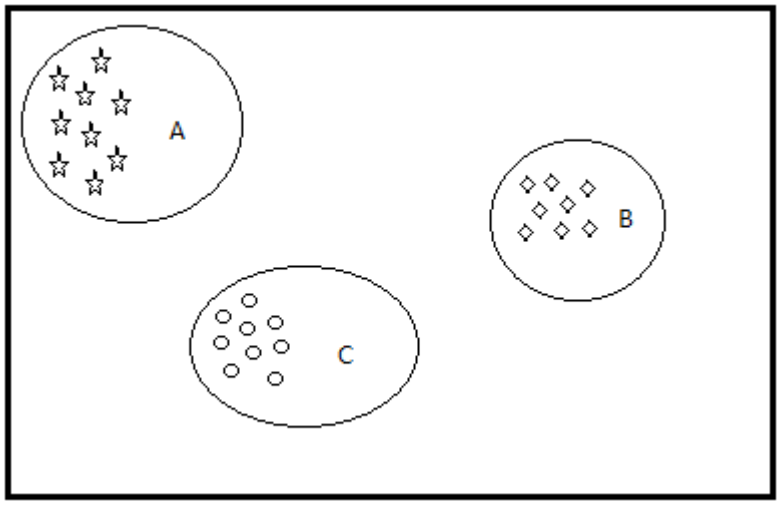


Figure 11 Results of Clustering Analysis

After we understand the meaning of cluster analysis, the next question is, why is it useful in QA system?

In QA system of open domain, especially in enterprise, there are a large number of texts stored in the knowledge databases. In QA system, we call these texts questions and answers. With time going by, along with the updating, the amount of data will be increased dramatically. As the result, the phenomena of information overload are often occurred.

To deal with this problem, text clustering is one of the most important approaches. Through cluster analysis, the orderless texts can be handled by different kinds of feature. It is more convenient for users to search information from the database. As the results, information retrieval will speed up immediately, besides, the cost will be decreased during the maintenance, additionally, if you do the retrieval in one cluster, the accuracy will be higher because of the small amount of data, compared with the big total group.

## 6.1.2. Classical K-means Algorithm

One famous algorithm of clustering analysis is K-means algorithm; here we give the definition of K-means algorithm.

K-Means is a simple but famous algorithm for cluster analysis. Again all objects need to be represented as a set of numerical features. In addition the user has to specify the number of groups (referred to as k) he wishes to identify.

Every object is represented by some feature vector in an n dimensional space, n is the number of all features used to describe the objects to cluster. Then k points are chosen randomly in that vector space, these K points serve as the initial centers of the clusters. After it, all objects are assigned to the center they are closest to. Usually users or the target task determine the distance measure.

Next step, for every cluster, there is a new center; every center is computed by averaging the feature vectors of all objects assigned to it. Then repeat the procedure of assigning objects and re-computing centers until the process converges. The algorithm can be proven to converge after a finite number of iterations.

Now I will show the standard K-means algorithm.
In this part, the standard K-means algorithm will be briefly described. The number of clusters k is assumed to be fixed in k-means clustering. Let the k prototypes $(w_1, \dots, w_k)$ be initialized to one of the n input patterns $(i_1, \dots, i_n)$. Therefore,

$$w_j = i_l, j \in \{1, \dots, k\}, \mathrm{l} \in \{1, \dots, n\}$$

Below we will show a high level description of the k-means clustering algorithm [1]. $C_j$ is the $j^{th}$ cluster whose value is a disjoint subset of input patterns. The quality of the clustering is determined by the following error function:

$$\mathrm{E} = \sum_{j=1}^{k} \sum_{i_l \in C_j} \left| i_l - w_j \right|^2$$

Function k-means()
Initialize k prototypes ($w_1, \dots, w_k$) such that $w_j = i_l, j \in \{1, \dots, k\}, l \in \{1, \dots, n\}$
Each cluster $C_j$ is associated with prototype $w_j$
Repeat
  for each input vector $i_l$, where $l \in \{1, \dots, n\}$,
    do
      Assign $i_l$ to the cluster $C_{j*}$ with nearest prototype $w_{j*}$
  for each cluster $C_j$, where $j \in \{1, \dots, k\}$, do
      Update the prototype $w_j$ to be the centroid of all samples currently in $C_j$, so that
      $w_j = \sum_{i_l \in C_l} i_l / c_l$
  Compute the error function:

$$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} |i_l - w_j|^2$$

Until E does not change significantly or cluster membership no longer changes

The appropriate choice of k is problem and domain dependent and generally a user tries several values of k. Assuming that there are n patterns, each of dimension d, the computational cost of a k-means algorithm per iteration (of the repeat loop) can be decomposed into three parts (Saida, Srinivas, & Sivaram, 2012):

1. The time required for the first for loop is $O(nkd)$.
2. The time required for calculating the centroids (second for loop) is $O(nd)$.
3. The time required for calculating the error function is $O(nd)$.

## 6.1.3. Traditional Ant-tree Algorithm

The Ant-Tree algorithm is based on the self-assembly behavior. The behavior is observed of ants. The ants live in certain species, and the species structures are used as bridges to build the nest. The structure is built by using an incremental process in which ants joint a fixed support or another ant for assembling. Ant-Tree builds a tree structure representing a hierarchical data organization which divides the whole data set. Each ant represents a single datum from the data set and it moves in the structure according to its similarity to the other ants already connected to the tree under construction (Errecalde & Ingaramo, 2010).

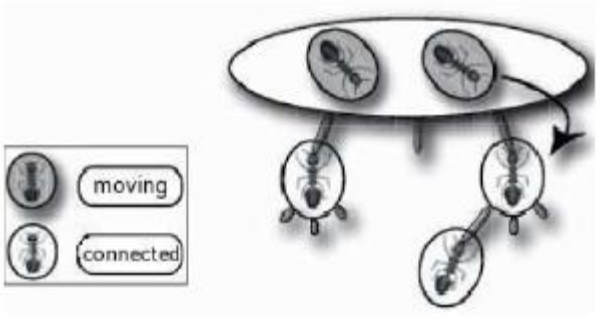


Figure 12 Tree structure generation by self-assembling artificial ants (based on Errecalde & Ingaramo, 2010)

Every node in the tree structure represents a single ant and each ant represents a single datum. The main aspect in structure is the decision where each ant will be connected, either to the main support or to another ant.

Each ant to be connected to the tree represents a data to be classified. Starting from an artificial support called $a_0$, all the ants will be incrementally connected either to that support or to other already connected ants. This process continues until all ants are connected to the structure. Each ant $a_i$ has associated the following terms:

1. $\tau(a_i)$, the ingoing links of $a_i$. A set of links toward $a_i$ (the $a_i$'s children).
2. $O(a_i)$, the outgoing link of $a_i$. A link to its parent node (the support or another ant).
3. A datum $d_i$ represented by $a_i$.
4. Two metrics called respectively similarity threshold ($T_{Sim}(a_i)$) and dissimilarity threshold ($T_{Dis}(a_i)$) which will be locally updated during the process of building the tree structure.

Figure 11 shows a general outline of the self-assembling of artificial ants. It can be observed that each ant $a_i$ is either of the two following situations:

1. Moving on the tree: a walking ant $a_i$ can be either on the support ($a_0$) or on another ant ($a_{pos}$). In both cases, $a_i$ is not connected to the structure. Consequently, it will be free of moving to the closest neighbors connected to either $a_0$ or $a_{pos}$. In Figure 4 is showed the neighborhood corresponding to an arbitrary ant $a_{pos}$.
2. Connected to the tree: in this case $a_i$ has already assigned a value for $O(a_i)$, therefore, it cannot move anymore. Additionally, an ant is not able to have more than $L_{max}$ ingoing links ($|\tau(a_i)| \leq L_{max}$). The objective is to bind the maximum number of incoming links.

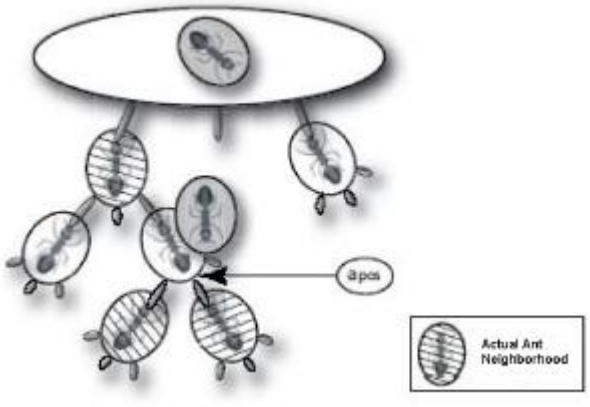


Figure 13 Neighborhood corresponding to an arbitrary ant $a_{pos}$ (based on Errecalde & Ingaramo, 2010)

Algorithm is shown below:

```
Let τ be a list (possibly sorted) of ants to be connected
Initialize: Allocate all ants on the support.
TSim(aj )←1 and TDissim(aj )←0, for all ant aj
Repeat
1. Select an ant ai from list τ
2. If ai is on the support (a0)
then support case
else ant case
Until all the ants are connected to the tree
```

## 6.1.4. New Ant-Tree Algorithm

In this QA system, we improve the traditional cluster analysis algorithm, namely, we combine the K-means algorithm and ant-tree algorithm together. Our new algorithm is the ant-tree cluster analysis based on K-means algorithm. The main improvement will be discussed in this section.

To make sure the initial sequence of the ant which connect to the root; we hope that the ant which connect to the root will close to the cluster core. We put forward the concept of the silhouette coefficient. The silhouette coefficient combines the degree of similarity and dissimilarity. For individual:

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$
$$(-1 \le S_i \le 1)$$

$\alpha_i$: The Average dissimilarity between di and the individual in the same group
$b_i$: The Average dissimilarity between di and the individual in the nearest group
For global:

$$S_k = \frac{1}{n}\sum_{i=1}^{n} S_i$$

k: The number of clustered group.
$S_k$: Silhouette coefficient

The new algorithm:

```
C = K-means (P: NodeSet, K:Int); // C is a clustering of P
Calculate array's a and b;
S = array(K);
foreach (C as Cluster) {
 S[i] = (b[i] - a[i]) / max(a[i],b[i])
}

foreach (C as Cluster) {
  Ant ant = takefirst (Cluster);
  Tree antTree = Tree (ant);
  while ( ! empty(Cluster) ) {
node = takenext (cluster);
process node into antTree;
  }
}
```

Evaluate average silhouette coefficient:

1. K-means algorithm to do the clustering to get the initial clustered group, the number of the groups is K
2. Set up an array to hold the cluster values, so the size of this array is K elements. Each element of this array corresponds to some clustered group, to store the every individual silhouette coefficient of every group, Descending
3. Take the first ant from each clustered group, and transform this into a single-node ant tree.
4. Iteration, until the entire array become empty.

## 6.1.5. Similarity Based on Hownet Knowledge Base

Traditionally, the similarity of words:Sim($W_1, W_2$) has the relationship with the distance of words Dis($W_1, W_2$)
$W_1, W_2$: words 1 and 2

$$\mathrm{Sim}(W_1, W_2) = \frac{a}{Dis(W_1, W_2) + a}$$

$a$: Adjustable parameters

Hereby we use the algorithm based on Hownet database to calculate the similarity after clustering:

Firstly, we introduce some definitions:

Similarity between words

$W_1$: Words

$W_{1i}$: contains n concepts, $S_{11} \ldots S_{1n}$

$W_{2j}$: contains m concepts, $S_{21} \ldots S_{2m}$

$$\mathrm{Sim}(W_1, W_2) = \max_{i=1..n, j=1...m} Sim(W_{1i}, W_{2j})$$

Similarity between sentences:

Average weights averaging algorithm:

Set up the matrix $R_{(m,n)}$: Sentence A has m words, sentence B has n words.

Similarity between A and B: the sum of the similarity between each two words from matrix.

$$S = \sum_{i=1,j=1}^{m,n} R_{ij}(mn)$$

Text feature extraction:

Text t contains p sentences.

Set up the matrix RPP like R,

$$M_p = \sum_{i=1}^{p} S_{pi}$$

Finally, we choice the highest value of MP to be the text feature.

## 6.1.6. Results of Similarity Test

Now we can use this algorithm to test some sentence and paragraphs:

Table 4 Result of Calculating Similiarity

| R | Beijing | Shanghai | Are | Two | Big | City |
|---|---|---|---|---|---|---|
| **They** | 0.04444 | 0.04444 | 0.074074 | 0.074074 | 0.04444 | 0.04444 |
| **Both** | 0.07333 | 0.07333 | 0.068254 | 0.048971 | 0.86111 | 0.8 |
| **Belong** | 0.04444 | 0.04444 | 0.285714 | 0.074074 | 0.04444 | 0.04444 |
| **China** | 0.90741 | 0.90741 | 0.655026 | 0.047462 | 0.65503 | 0.64074 |

| S | Beijing and Shanghai are two | Beijing is in the north and | Shanghai | They both belong to China |
|---|---|---|---|---|

| | big cities | is in the south | |
|---|---|---|---|
| **Beijing and Shanghai are two big cities** | 1 | 0.262653 | 0.32505 |
| **Beijing is in the north and Shanghai is in the south** | 0.262653 | 1 | 0.218664 |
| **They both belong to China** | 0.32505 | 0.218664 | 1 |

| | Beijing and Shanghai are two big cities | Beijing is in the north and Shanghai is in the south | They both belong to China |
|---|---|---|---|
| **MP** | 1.587703 | 1.481517 | 1.543714 |

## 6.2 Key Words Extension

### 6.2.1. Why to Do the Key Words Extension

In the question-answering sentences, some words are not usually the key words of the original questions, but extended synonymous of these words. For example, the problem is “What color are tomatoes?” The sentence of the answer is "tomatoes are red". The question used “Tomatoes" while the answer used the word “Tomato". This creates a fail of keyword query. Hence we need an appropriate extended of keywords.

Keyword expansion of the system increases the recall rate, but not appropriate will greatly reduce the retrieval accuracy, so general question answering system is a very cautious about keyword expansion. In the development of the system, the extending of keyword will be completed in the following three areas.

1. Key word synonym Expansion: use all synonyms of keywords as extended keywords, using Semantic Dictionary and Knowledge Network to complete keyword synonym expansion.
2. Expansion According to the type of answer: For some types of questions, the corresponding answers are always containing some common features about words. For example, to ask the location, the answer often appears in “at", " lie ",” in the ", then expand the words as keywords.
3. Key words expansion of the upper and lower: In order to improve the rate of retrieval recall, sometimes upper words of keywords as a supplement to join Keywords. For example, the country name “Finland "appears in problem, then the “European” will join key words.

### 6.2.2. How to Do the Key Words Extension

We assume that, the number of key words in question is m, among of them, the number of the extensible key words is n, t is positive integer. The initial value of t is 0. Absolutely, $m \geq n \geq 0$.

According the question of input, we compute the value of m and n through question analysis and synonyms inquiry.

---

Algorithm:
{
If m=0
{
Return;
}
Set t=0;
If $\mathbf{t} \leq \frac{n}{2}$
{
Choice t key words from the n extensible key words to replace all the $C_n^t$ displacement
{
Search the key words set after displacement as the initial order, also set up the candidate answer documents conclude both the $\frac{10\times(m+1-t)}{C_n^t}$ downloading retrieval results from search engine and downloaded abstract documents;
}
t=t+1;
}
Return;
}

---

## 6.3 Answer Extraction:

### 6.3.1. Traditional Algorithm:

Vector space model (VSM) is a method applied most frequently, with a relatively better result. In 1969, Gerard Salton put forward vector space model, which is a file representing statistic model. The model's core idea is to mapping each file into a point in vector space constituted by a set of regulated orthogonal entry vector. As for all files and unknown files, entry vector

in space can be introduced, for example T1,W1,T2,W2,…,Tn, Wn, with Ti as featured vector entry, and Wi as Ti's weighting. It usually needs to construct an evaluation function to represent entry weighting, with the only calculating rule as differentiate files to its largest extent. The advantage of such vector space model is that unstructured texts and structured texts can be described in vector form, making it under mathematical process possible.

In respect of information retrieving, applied formula in practice is:
Answer A represents a vector, n represents amount of words, aj is the frequent of word No.j in answers. In a similar way, question Q can represent a vector q. The degree of correlation between question and answer can be obtained by calculating cosine of two vectors.

After applying into some examples by tests, the result is not that obvious. Only sentences consist of many words can this method exhibits a sound result, because it's just a statistic method. The more words contained in a sentence, the more frequently the related word will appear. Therefore, the effect of such statistic method can be reflected. However, in our system, what we face is single sentence, which has the number of words less than enough to reflect such sound effect.

Vector space model fails to take full advantage of other useful information in questions and answers. For example, the order of keywords, distance between keywords, and length of questions and answers, etc., which exerting a significant influence on extracting answers. Therefore, we put forward a method using sentence similarity to calculate similarity degree between question and answer.

VSM (Vector Space Model)

$$\text{cosine}(a,q) = \frac{\vec{a}\cdot\vec{q}}{\left|\vec{a}\right|\left|\vec{q}\right|} = \frac{\sum_{i=1}^{n} a_i \times q_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \times \sqrt{\sum_{i=1}^{n} q_i^2}}$$

$$a = (a_1, a_2, \dots, a_n)$$

## 6.3.2. New Algorithm

Since the backwards of VSM algorithm, we take the factors into consideration:
The order of keywords, distance between keywords, and length of questions and answers. Now I will show the details of new advanced-VSM algorithm:

Answer: A (vector)
n: number of words
$a_j$: times which words j appears in A.
Q: question (vector)
Cosine: the Relevance between question and answer.

Example:
Question: Who is the Chinese Outer Space's first man?

Answer: Yangliwei, the first Chinese man went to Outer Space.

Now we divide the two sentences into words and catch up with the key words:
Q: Chinese; Outer Space; First; Man.
A: Yangliwei; First; Chinese; Man; Went; Outer Space.

Factor:
The number of the same key words (the bigger, the better)
KeyWC(Q): The number of Non-repetitive key words in question.
KeyWC(A): The number of Non-repetitive key words in answer which appear in question.

$$\mathrm{WordSim}(A, Q) = \frac{\mathrm{KeyWC}(A)}{\mathrm{KeyWC}(Q)}$$

Example: KeyWC(A)=4; KeyWC(Q)=4;
WordSim(A,Q)=1.
The length of sentence (the bigger, the better)
Len(A): the length of answer, or the number of words in answer.
Len(Q): the length of question, or the number of words in question.

$$\mathrm{LenSim}(A, Q) = 1 - \frac{|\mathrm{Len}(A) - \mathrm{Len}(Q)|}{\mathrm{Len}(A) + \mathrm{Len}(Q)}$$

Example: Len(A)=6; Len(Q)=7;
LenSim(A,Q)=0.923.
The order of key words in sentence (the bigger, the better)
Seq(A): the order of natural number that the key words of answer which appear in question.
Rev(A,Q): the reversed number of Seq(A)
MaxRev(A,Q): the maximum reversed number of Seq(A)

$$\mathrm{OrdSim}(A, Q) = 1 - \frac{\mathrm{Rev}(A, Q)}{\mathrm{MaxRev}(A, Q)}$$

Example: Rev(A,Q)=3; (order: 5,3,6,4)
MaxRev(A,Q)=6; (order: 4,3,2,1)
OrdSim(A,Q)=0.5
The distance between key words (the smaller, the better)
Dis(A): the distance between key words that appear in the question
Dis(Q): the distance between the most left key word and most right key words in question

$$\mathrm{DisSim}(A, Q) = 1 - \frac{\mathrm{Dis}(A)}{\mathrm{Dis}(A) + \mathrm{Dis}(Q)}$$

Example: Dis(A)=4
Dis(Q)=3
Dis(A,Q)=0.428

$$\mathrm{Sim}(A, Q) = \lambda_1 \mathrm{WordSim}(A, Q) + \lambda_2 \mathrm{LenSim}(A, Q) + \lambda_3 \mathrm{OrdSim}(A, Q) + \lambda_4 \mathrm{DisSim}(A, Q)$$
$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$$

# 7. Testing of QA System

## 7.1 Methods of Testing

In a question-answering (QA) system processing natural language, technical evaluation plays an important role in advancing the development of technology. In a typical technical evaluation, the organizer presents identical data, methods and standards, through which participants compare and contrast with each other. Through comparison and contrast, participants recognize their own advantages and disadvantages. Hence, research can progress.

Some major international evaluations and tests for specific domains will be briefed on next.
Since the 1980s, burgeoning development of information extraction research has been witnessed, thanks to the promotion of MUC short for Message Understanding Conference (Ralph Grishman & Sundheim, 1996). From 1987 to 1998, MUC, sponsored by the Defense Advanced Research Projects Agency (DARPA), was held seven times. Besides MUC, America's National Institution of Standard Technology (NIST) performs Automatic Content Evaluation (ACE) (Strassel et al., 2008).

Since 2002, NIST has been annually holding a machine translation evaluation, which is part of its TIDES project. The evaluation target is a machine translation system which is based on rules or statistics. Such evaluation has become a major event.

The most influential international information retrieval evaluation and test is the annual Text Retrieval Conference (TREC) jointly organized by NIST and DARPA. The TREC evaluation and test commencing since 1992 has been attracting many organizations to participate. It mirrors the global development of information retrieval technology. Now, two sub-evaluation and sub-test items have derived from TREC, namely, TRECVid (a video-based evaluation) and QA (Question-Answering).

The QA system in my thesis is adapted to the TREC QA evaluation. The following is the description of TREC research objectives and contents.

A. Research objectives
1) To promote the retrieval research based on large literature sets. There must be sufficient experiment corpuses to make sure a system possesses diverse themes. In this case, the capacity of TREC literature set usually reaches 2G, in other words, its collected literature can amount to a range from 500,000 to 1,000,000 pieces.
2) To establish an open forum to facilitate the exchanges between enterprises, academic organizations and government departments
3) To verify the validity of retrieval in solving concrete problems to expedite the commercialization of experiment technologies. In case a certain technology demonstrates satisfactory experiment results, it can be rapidly commercialized

4) To provide large corpuses, unified test procedures to sort out test and evaluation results systematically to seek better text retrieval evaluation and test methods, and to offer enterprises and academic organizations access to proper retrieval evaluation technologies, and to develop new evaluation technologies better corresponding to the current system.

B. Research contents

In a typical TREC, a variety of research groups share a data source, complete the same retrieval tasks so as to compare and contrast various systems and methods. Therefore, TREC consists of three sections, namely, literature set, retrieval mode set, and correlation identification. To make sure a system has a wide range of themes, a literature set must give holistic regard to the interrelated aspects of selection of words and phrases, literature styles, and literature formats. TREC can perfectly meet such demand because it enjoys a huge reserve of literatures, whose sources mainly come from newspapers, journals and governmental literatures (such as reports from Department of Energy, and patents). Each document is marked with SGML short for Standard Generalized Makeup Language and given a unique document Number (DOCNO). Documents are kept intact so that they can maintain their consistency. For example, the spelling errors of a document are kept. Besides, the collected documents can be either long or short, either full or just partial. It stands for the diversity of the data base with which we are retrieving.

As TREC evaluation is based on a Granfield paradigm, the effectiveness of a system is assessed in line with the number of relevant documents retrieved. TREC evaluation is used to reflect recall ratio and precision ratio. In most TREC programs, two items are combined into one performance index called Mean Average Precision (MAP). Precision ratio is measured to a given question mode when a relevant document is retrieved. Accordingly, MAP refers to the mean value of all retrieval question modes.

In addition, some TREC projects need different evaluation units. For example, a QA project aims at finding the most optimal answer from all the answers among the output sequence to a certain question, that's, the output that is ranked first. The evaluation unit is Mean Reciprocal Rank (MRR).

## 7.2 Evaluation

In this section, we will evaluate the proposed QA system from the following aspects.

(1) The experiments of verifying accuracy about problem classification

In the stage of understanding problem, if your analysis about problem type is inaccurate, the answer extracted will be not correct. So first of all, to test the accuracy of the type of problem identification, and compared with the general classification methods of problem. The general problem of classification methods is to extract the question word in question and get the type of problem identification results.

(2) The experiments of testing the extended results of key words

The experiment is to complete the simple word segmentation and the extraction of keywords about

general question, then retrieve of relevant documents; in addition, complete keyword expansion, and you will get results with Google Search.
The way 1: Use Google Search after word segmentation
The way 2: Use Google Search after word segmentation and extension
Test Question 1: What is the color of tomatoes?
Key words: tomato color
Extended Key words: tomato color tomatoes
Test Question 2: Why have a sandstorm phenomenon?
Keywords: sandstorm phenomenon
Extended Key words: Causes of sandstorm phenomenon

Method 1 is just using a collection of key words to search for an answer, almost intelligent search engines use the method today, then ignoring the diversity of expressions of Chinese. Method 2 not only includes method 1 but also extends the question on the basis of maintaining the original semantic, and then it will be in line with the Chinese expression Habits and widen the range of valid search of the answer. Apparently, after keyword expansion, the retrieval accuracy increase effectively, this method not only controls the scope of the search results, and avoids the emergence of a lot of irrelevant search results.

(3) The test of semantic role labeling results
The test of semantic role labeling results uses 3 commonly used tests of evaluation indicators, Precision P (precision), Call Back to the rate R (Recall ) and f-measure. Calculations are as follows:
P= Mark the correct semantic blocks, mark semantic total number of blocks
R= Mark the correct semantic blocks, test total number of corpus and semantic of blocks

$$F_\beta = (\beta_2 + 1) \times P \times R / (\beta_2 \times P + R)$$

Here, $\beta = 1$

After analysis and statistics, we will get some factors affecting the accuracy of the system, they are not related to the performance of the system itself, but it is impossible to avoid.

(1) Input errors of user
Wrongly written or mispronounced characters you enter as the user may occur and cause the system cannot analyze the problem.

(2) Online Search
Problem of network search analysis algorithm occurred mainly in the area of Google search engine, after searching some keywords, the keywords on the feedback page have not been marked in red, so it causes the consequence about that Named Entity Recognition Module could not be identified.

# 8. Conclusion

## 8.1 Conclusion of Current Research

This paper emphasizes on the description of prototype frame of QA system, and three new algorithms of each component in the system frame. In addition, it provides deeper insight and analysis of question comprehension and information retrieval in QA system. Moreover, the methods proposed in this thesis are proved to be reasonable, practical and theoretical by testing parts of prototype system. Following work have been done in this thesis:

(1) This thesis puts forward the concept of "HowNet", which is common sense knowledge base with concepts represented by Chinese and English words respectively as descriptive objects and relationships between concept and concept, and between properties of concept. Compared with "WordNet", "HowNet" can solve problems between lines of Chinese text better. The major contribution of "HowNet" lies in a new comparison of words similarity. In "HowNet", each concept won't absolutely match to a certain object in the concept hierarchy with tree structure. Instead, each concept is described with language of a certain field of knowledge by a series of resources, which in turn form a resource hierarchy with tree structure through hyponymy. This thesis adopts calculation of word similarity based on "HowNet", and calculates similarity of two semantic expressions described in such knowledge language. Proved by experiments, the new method can effectively extract textual features.

(2) This thesis improves Ant-Tree method applied for short text clustering, in order to improve the performance of clustering. Based on Ant-Tree method, this thesis combines K-means method, introduces profile coefficient, adjusts initialization of Ant-Tree method, and enhances clustering quality. Proved by experiments, this method increases efficiency of clustering and texts searching and refines the clustering precision as well.

(3) In respect of issue categorization, taking the tremendous workload of analyzing syntactic and semantics into account, it differs from the traditional machine learning method in normal QA system. Instead, it takes advantage of recognizing interrogative pronoun and essential questions, with revised method of heuristic rules, to ensure efficiency of the whole system. As for questions based on fact or questions which have simple grammar, such method provides a better answer.

(4) In order to increase retrieving efficiency of QA system, the framework establishes FAQ database which memorizes frequently asked questions and their corresponding answers.

(5) In regards of system structure, this thesis compensates for the shortage of traditional QA system in the field of consumer feedback, and provides feedback functions. In addition, due to the establishment of QA database for memorizing feedback results, it enables the system

ability of learning, and effectively increases user searching speed and accuracy rate.

(6) Taking advantage of online knowledge sharing platform, the framework constructs a large-scale authentic user QA database, realizes the combination of local knowledge base indexing and search engine indexing, and increases response speed of system and accuracy rate of answers.

(7) Compared to the traditional QA system, the dynamic QA system constructed in this thesis can update knowledge database in accordance with user requirement to maintain searching accuracy. At the same time, in accordance with user requirement, it can delete redundant information and outdate information to ensure the scale of system stay among affordable range and decrease risk of system failure and costs of maintenance.

## 8.2 Further Research

There is still a gap between QA prototype system designed in this thesis and the actual QA system. In addition, theories and methods put forward in this thesis still need to improve. Therefore, following issues are proposed for further research.

(1) Question classification based on rules
In respect of question classification, although the combination of enlightenment-oriented interrogative and focus-oriented question can increase the speed of searching answers, in an open field circumstance, the types and structures of question are complicated. The classification method proposed in this thesis can only solve partial problems. Difficulties are aroused in the fields of question classification. Up till now, the QA system described in this thesis is basically designed for matters of fact which have brief and short answers. Further research will focus on how to refine the types of questions to provide solution in open field circumstance.

(3) Deeper insight of named entity
At present, this system only extracts simple named entity. And the further research into extracting complicated entity will be developed. The following major task is to consider how to introduce information extraction technique of complicated entity into the QA system, for example, automatically summarizing technology.

(4)Exploration of answer extraction technology
This thesis mainly emphasizes on the initial two phases of QA system, and covers a little on the research into answer extraction technology. The semantic similarity calculation method put forward in this thesis can satisfy the requirement of answer extraction to a certain extent, but the semantic framework can only solve superficial semantic match issues. Therefore, the further study will start with deeper semantic analysis to gradually increase accuracy of answer extraction.

All in all, QA system is a new generation of search engine integrated with natural language

processing technology and information retrieval technique. With continuous research in this field, the study of QA system in future will make great breakthrough of traditional design methods of QA systems. This thesis contributes to the development of the next generation QA system both theoretically and practically, and some enlightenment for further research.

# Reference


Abney, S. (1997). Part-of-Speech Tagging and Partial Parsing. In S. Young & G. Bloothooft (Eds.), *Corpus-Based Methods in Language and Speech Processing* (pp. 118-136): Springer.

Apache. (2010, 2010). Welcome to Apache OpenNLP. 2013, from http://opennlp.apache.org/

Buyko, E., Wermter, J., Poprat, M., & Hahn, U. (2006). *Automatically Adapting an NLP Core Engine to the Biology Domain*. Paper presented at the the Joint BioLINK-Bio-Ontologies Meeting: A Joint Meeting of the ISMB Special Interest Group on Bio-Ontologies and the BioLINK Special Interest Group on Text Data M ining in Association with ISMB, Fortaleza, Brazil.

Charoenpornsawat, P., Sornlertlamvanich, V., & Charoenporn, T. (2002). *Improving Translation Quality of Rule-based Machine Translation*. Paper presented at the the 2002 COLING workshop on Machine translation in Asia.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51-89.

Dong, Z., Dong, Q., & Hao, C. (2010). *HowNet and its computation of meaning*. Paper presented at the the 23rd International Conference on Computational Linguistics: Demonstrations, Beijing, China.

Errecalde, M. L., & Ingaramo, D. A. (2010). A new AntTree-based algorithm for clustering short-text corpora. *Journal of Computer Science & Technology, 10*(1).

Lashkari, A. H., Mahdavi, F., & Ghomi, V. (2009). *A Boolean Model in Information Retrieval for Search Engines*. Paper presented at the International Conference on Information Management and Engineering 2009.

Liddy, E. D. (2009). *Natural Language Processing for Information Retrieval* (Vol. null): Taylor & Francis.

Molina, A., & Pla, F. (2002). Shallow Parsing using Specialized HMMs. *Journal of Machine Learning Research, 2*, 595-613.

Ralph Grishman, & Sundheim, B. (1996). *Message Understanding Conference - 6: A Brief History*. Paper presented at the the 16th International Conference on Computational Linguistics, Copenhagen, Denmark.

Roberts, I., & Gaizauskas, R. (2004). *Evaluating Passage Retrieval Approaches for Question Answering*. Paper presented at the the 26th European Conference on Information Retrieval.

Saida, S. J., Srinivas, L., & Sivaram, R. (2012). An Efficient K-Means and C-Means Clustering Algorithm for Image Segmentation. *International Journal of Science and Applied Information Technology, 1*(3), 84-87.

Strassel, S., Przybocki, M., Peterson, K., Song, Z., & Maeda, K. (2008). *Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction*. Paper presented at the the 6th International Conference on Language Resources and Evaluation.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*: Addison-Wesley.

Waltz, D. L. (1978). An English language question answering system for a large relational database. *Communications of the ACM, 21*(7), 526-539. doi: 10.1145/359545.359550