Implementation and Evaluation of a Question Answering system using Concept-based Vector Space Model

Israrni Ismail Takashi Yukawa
Department of Electrical and Electronical System
Nagaoka University of Technology
1603-1 Kamitomioka - machi, Nagaoka, 940-2188 JAPAN
israrni@stn.nagaokaut.ac.jp yukawa@vos.nagaokaut.ac.jp

***Abstract*— In this paper, a Question Answering (QA system) system using the concept based vector space model is proposed, and practically constructed and evaluated. The concept-based model is a model based on the consideration of the human language. Since the concept-based model is providing the relation of similarity among the words, the retrieval of the target document with the consideration of its significant meaning becomes possible. The QA system is an information retrieval system that works on natural language queries and returns a short phrase or a word as the answer. The system is commonly composed of the retrieval of the related document and the determination of the answer word from the extracted documents. However, the miss-dropped of the significant documents and the difficulty of identifying the answer word in the case of common nouns being expected as the answer words, are still unsolvable in the present method. As a solution, the retrieval and words are extracted with the consideration of their semantic are proposed by using the concept-based model.**

***Index Terms*— concept-based vector space model, question answering system.**

I. Introduction

Recently, as the usage of computers had become a part of human life, the simplicity in handling computers has become much more important. For many years, a human user needed to learn how to communicate with computer system before he could handle the system. Meaning that only those who have expert knowledge of the machine language are able to communicate (operate) with the computers. In accordance with the progress in the information technology, the usage of computers has been simplified so that beside the experts, others are able to operate computers too.

In that sense, a new era is coming in the communication method between the human and computer system. In the near future, human user (common user) have less need to learn certain computing language in order to operate the computer, but conversely the computer will need to understand human needs from natural language expression.

Consequently, research on the knowledge extraction of information written in natural language had become one of the hot topics. This paper focuses on the question answering system using the Concept Based Vector Space Model as one of the natural language based knowledge extraction systems. The system that will be introduced in this paper focuses on the ability of answering the question searching for the common nouns answer.

II. Concept based vector space model (CBVSM) as a method of natural language recognition

The recognition of natural language can be performed using the combination of singular expression and rule-based system but the formation of rules take a lot of time and are costly. One of the new approaches for natural languages recognition is the Concept-based Vector Space Model (CBVSM).

The concept base model is knowledge base of the words contained in the document set. The idea of CBVSM is that semantic words appear in a similar way and each word's concept can be recognized by the way they appear in the sentences. Vectors are assigned to represent each word. The vectors of each word are generated by calculating the co-occurrences of the words composed in the target document. In CBVSM similar words will possess a large cosine co-efficient between.

Here, the CBVSM is proposed as a new approach for question answering system. Question answering system (QA system) is a kind of retrieval system that works on a natural language query and returns the answer in short phrase or words. A QA system is composed of two important parts; the retrieval of related documents and the extraction of the answer word from the related documents. The details on these two sections will be discussed later in this paper.

A. Generation process of the concept base

In the concept-based model, as the words vector are generated using the target documents, the words' co-occurrences in the target documents are statistically calculated. There are several ways proposed to represent a word's conceptual knowledge and the most common method is by representing the word as a group of attributes and its value. However, at the present technology level, the automatic acquisition of the word concept still remains difficult. Accordingly, a simple concept-base definition is proposed. The concept, $Word_i$ for the word i located in the concept-base can be express as;

$$Word_i = \{(p_i1 , q_i1),\ldots,(p_in , q_in)\}$$

The characteristics of the concepts can be obtained from the independent root words among the documents. Moreover, for the attribute with high appearance frequency within the sentences, the word's concept can be considered as distinctive, that appearance frequency of 10 words after and before the attribute should be taken as weight. For example, if the explanation for the [Nagano Olympic];

...20th century... last.. winter ...Nagano-Olympic
...started 20 January ... Nagano prefecture

10 words

The concept knowledge for the word [Nagano Olympic] can be observed as in Table 1;

Table 1: the co-occurrence matrix of words

	20	century	Last	winter	Nagano olympic	started	January	…
20								
Century								
Last								
Winter								
Nagano Olympic	2	1	1	1		1	1	
Started								
January								

By calculating the co-occurrence of each all words in the target documents, the co-occurrence matrix of the words will be generated.

Here, due to the limit of the computing amount resources, the words contained in the concept base is limited to 10,000 words. The words are chosen with priority on the words with high co-occurrences frequencies. The words with lower co-occurrences frequencies in the document collection will have a higher probability of being dropped out from the concept base.

Consequently, the dimension of the co-occurrence matrix is reduced to 100,000×3,000 dimensions (as in table 2). The specification of the matrix is indispensable in order to calculate the concept-based of 10,000 words. It is due to the limit of the computing amount to do the SVD on the co-occurrence matrix. However, although the matrix is set this way, the effect to the retrieval performance is small, as the dropped axis are the words with lower co-occurrence frequency compared to the words selected to represent the 3,000 dimension of the matrix [1][2]

Table 2; the dimension reduction of the word's co-occurrence matrix

3000 dimensions

10,000 dimensions

	Word A	Word B	Word C	
Word A	0	3	1	
Word B	2	0	0	
Word C	0	2	0	
⋮				

This matrix then, compressed to 100~200 dimensions during the implementation of Singular Value Decomposition (SVD), finally results in generating the concept-based (Refer to table3). The compression of the dimension to 100~200 dimensions is based on the results experimentally proven before [1]. The results of the experiments show that the retrieval performance will run excellently at the range of 100~200 dimensions.

Table 3; the concept-based vector

100 dimensions

100,000 dimensions

	a1	a2	a3	
Word A	a11	a21	a31	
Word B	a12	a22	a32	
Word C	a13	a23	a33	
⋮				

B. Generation of the document base

The documents are defined by a set of words composed inside it. Document vectors, which are represented by each concept (words) vector composed in the particular document, are generated by the composition of the concept's vector within it.

$$Document_1 = \{Word_1 + Word_2 + Word_{3+} ...\} \cdots (1)$$

Combination of the word vectors will result the absolute value of the document vector to be bigger than 1. However, in order to make the calculation of the vector easier, the normalization of the document vector is executed before the

$$Document_{1'} = \frac{\{Word_1 + Word_2 + Word_3 + \cdots\}}{|\ Document_1\ |} \cdots (1')$$

retrieval process

From the definition, document vectors depend very much on the vectors of the words composed inside the document. This also means that the higher occurrence of a particular word in

the document will result in the document having a stronger relation to the word.

Before the generation of the document vector, the unnecessary words, known as stopwords should be removed. The stopwords, composed of the punctuation marks and words such as “は”, “が”, “です” etc, exists frequently in every document without representing the concept of certain documents.

C. Generation of the query vector

The query sentence supplied by the user is also a set of words. Therefore, the query can also be considered as a type of document. Query vector, in a similar manner to the document vector is represented by the word vectors composing a particular query sentence.

$$Query_{1'} = \frac{\{Word_1 + Word_2 + Word_3 + \cdots\}}{|Word_1 + Word_2 + Word_3 + \cdots|} \cdots (2)$$

In this case, a query vector will be generated every single time the user inputs a new query to the system.

D. Information retrieval of CBVSM

The information retrieval of the CBVSM model is executed based on the similarity degree between the documents and the query. Retrieval process is done by calculating the similarity degree of the query vector and the each document’s vector. The target documents then are sorted by the degree of their similarity against the query sentence. The most related document should be the document with highest similarity degree.

In the concept base, the query word, target documents and every word in the concept base are represented by vectors of the same dimension (refer figure1). This means that by calculating the similarity degree of these elements, the similarity between each of them can be identified. Given one query sentence, the most related document to the query can be calculated. The most related document will have the highest cosine co-efficient amongst the documents. Using the same approach, it is also possible to calculate the most related word to the given query.

$$\textbf{similarity:cos}\theta = \vec{d} \cdot \vec{q}$$

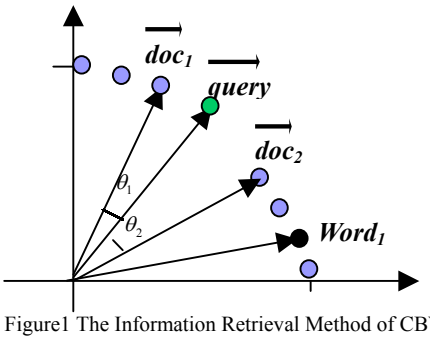

Figure1 The Information Retrieval Method of CBVSM

III. QA SYSTEM USING CBVSM

Figure2 shows the complete architecture of the proposed QA system using the Concept-based Vector Space Model (CBVSM-QAS).

In the CBVSM-QAS, the concept base and the document base are created using the original documents to be used in the retrieval module. In this paper, some improvements are proposed in the concept base generation module and the document base generation module in order to maximize the ability of the concept base during the retrieval of related documents. The details of the proposal will be discussed in section 4.

Whenever the user inputs a query into the system, the query sentence will be processed through the query vector generator. Every single word contained in the query sentence will be referred to the concept base. The words that are contained in the concept base then will be chosen as the query words (the

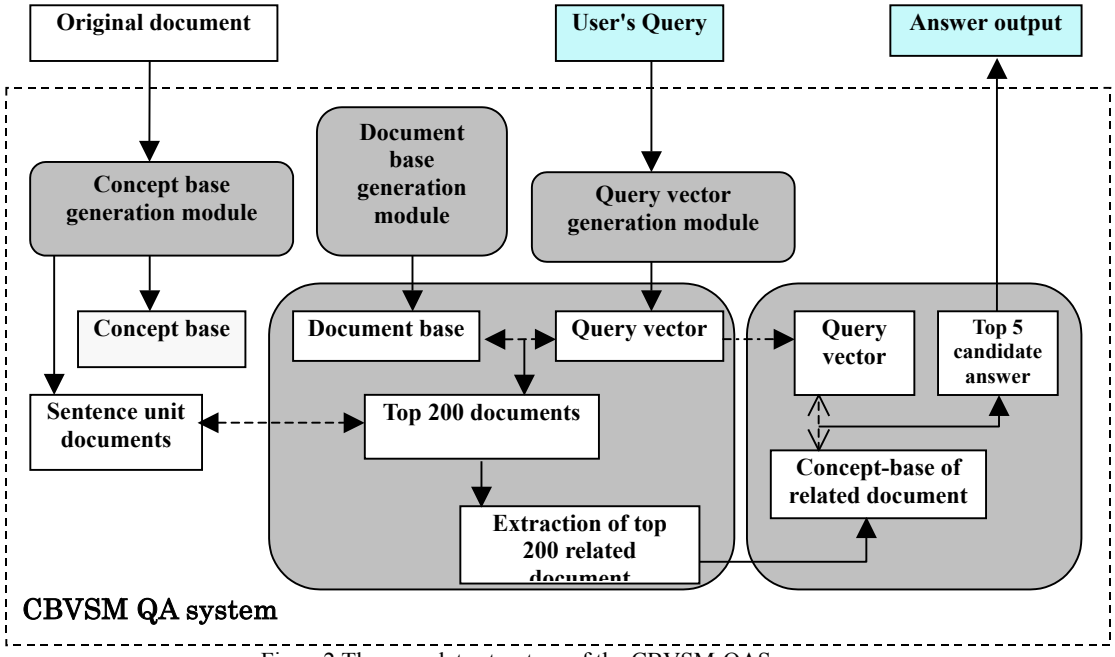

Figure2 The complete structure of the CBVSM-QAS

words to be searched). The composition of each chosen words' vector will generate the query vector.

The documents related to the query will then be search out from the similarity degree of this query vector and each document vector (document base). In this system the top 200 of the related documents that have been detected by the retrieval process will be extracted and stored in a different database. From the related document extracted, the answer words will be extracted in the answer word extraction module. The details of this part are is discussed in section 4.

After the extraction module is finished, the top 5 of the answer will be returned to the user.

IV. Related document retrieval method using CVBSM

In this section, the methods of the retrieval will be discussed in detail. Concept-based vector space model is proposed as the retrieval method of this system. It is expected that by using the concept-based vector space model, retrieval with consideration of the significance of the documents can be done. In order to ensure the best method of retrieval, the issue occurred during the implementation of the CBVSM into the question answering system and its solutions were verified. Moreover, the merit of the CBVSM comparing to the presently used model was confirmed.

In the effort of adopting CBVSM into the QA system, several tasks remain to be clean up in advance. Here we appoint these 3 issues;

1. The granularity issue on implementing the concept base into the QA system.
   In order to implement the concept-based model in QA system it is still not yet proven whether the direct retrieval of the whole document is possible or some modification is needed.
2. The compound words issues
   During the morpheme analysis, every compound word is splinted. The procedure will probably cause inappropriate vectors of the words to be generated in the concept base.
3. The elimination of the redundant words from the concept base
   Many of the redundant words to the QA system are contained in the concept-based. The elimination of these words should be done in advance to ensure that a more accurate concept-base can be generated.

A. Granularity issue on implementing the Concept Base into the QA system

In the original newspaper documents, the document lengths vary up to thousands of words in a document. In a certain document, some words only appear a few times, although the size of the entire document is big. Accordingly, as the generation of the document base depends a lot on the words frequency, the other words which have higher appearance frequencies will dominate the entire document vector. As a result, the concept of the words with low appearance frequency will not be emphasized within the document vector. This will probably result the document vector being relatively far from the word vector and making them less related to each other despite their semantic connection. In this situation, if the user makes a query on the words, the document which actually contains the words with small appearance frequencies but large in the document size will be considered to have less similarity to the query topic.

In order to confirm the issue, we tested 2 different methods. One uses the newspaper document as it was supplied into the CBVSM retrieval system. This type of retrieval method is named as the Document-unit CBVSM. On this Document-unit CBVSM, every newspaper document are used as a single target of retrieval process.The other method is by dividing each docurment into smaller documents. Using the original document supplied, the sub-documents are generated by splitting each sentence in the original one (later mention as Sentence Unit –CBVSM).

In the Sentence Unit CBVSM, the target document is replaced from the original documents to the short document collection created by dividing the original documents according to the sentences contained. By this method, most of the newly created documents will composed with average number of words and each word will fairly dominate the document vector.

B. Consideration of the compound words

During the generation of the concept base, the original document will be processed through the morphological analyzer. This process will divide the sentence into words. At the same time, each independent word will be stemmed to their origin making a set of the stemmed words generated after the morphological analysis process. This process is important in the generation of the concept base since it is important to clarify to the system the word that is going to be employed in the concept base. Furthermore, every word needs to be stemmed to it's origin in order to define the similar words. However, by dividing each word, some words that have different meaning when they appear together will also be separated.

Here, the proposal of combining the compound words on a rule base is proposed. In the proposed method, the regeneration of the compound words is executed after the morphological process. Based on the part of speech of the words, which are analyzed by the Chasen [3], the system combines the compound words into one word before expressing them as vectors.

In the conventional method, after the document processed through the morphological analyzer, every single word divided separately. For example, the words "year 2001"will be separated into "2", "0", "1" and "year". In the context of the present method, whenever the retrieval for document containing the word "year 2001", every single document containing the number "2", "0","1" and "year" will be considered as related. However, this phenomenon will surely effect on the precision of the retrieval performance, since the document which simply containing the words "2", "0", "1" and "year" comparing to the document containing the words "year 2001" is totally different. Consequently, the division

causes the words to be meaningless in the sense of document.

The re-generation of this kind of word is needed in order to improve the precision of the system's result. Here, the rules of the compound words' combination were studied. In the proposed method, some rules of combination of the words were contrived. For example;

1. (Numeral)*+ Numeral assist
2. Proper noun + Proper noun
3. Proper noun + Suffix

C. The elimination of the redundant words from the CBVSM

In the implementation of the CBVSM in the QA system, many of the words in the concept-based are indicated as inappropriate words to be appointed as the answer words. Words attribute by conjunction, adjective, postpositional particle, mark and adverb, exist frequently in the document. However, since most of the questions in the QA system are expecting nouns as the answer word, all of the other kinds of words will probably never be extracted as the answer to any question. Even in the case of verb like "検索する (searching)", after being process through the Chasen it will be outputted as " 検索(search)" and " する(do)". As the "検索(search)" stand as a noun, the word " する(do)" remain as a meaningless verb. Furthermore, the words as "する", is a very common word and used in almost of the sentence. This means that by letting the word stay as one of the elements in the document vector, the precision of the vector will be affected.

The generation of the Noun-CBVSM is based on the part of the sentence of the words contained in the document. After the morphological analysis process of the target documents (Sentence unit documents), the result of the analysis will be handed over to the Nouns Selection Module. The Nouns Selection Module will analyze on the part of the sentence of every word contained. In this process, only the nouns are being selected to remain in the concept base. The other type of words, since they are not considered to be important to the QA system will be ignored. As the result of this module, a new set of words which is composed of only the nouns will be produced and the generation of the concept base and the document base will start based on this result.

As a merit of this method, the manual elimination of stopwords before the generation of the concept-based (as mention in section 2), is able to be automated. In the current method, the system creator had to create a list of the stopwords. The elimination of the stopwords is executed by referring to the list created. However, the manual creation of the stopwords had always been inadequate, since it is quite hard for the creator to identify all of the stopwords used within the documents. Against this, by the generation of Noun-CBVSM, the elimination of the stopwords from the document can be done automatically since all of the stopwords are actually composed of the words beside the nouns.

D. The retrieval performance

The proposed method of improvement for the related document retrieval purpose is implemented and evaluated.

For comparison purpose, the Boolean model is employed. The Namazu system was used as a representative system of the Boolean model. The query words were chosen from the query sentence by eliminating the stopwords.

The result of this comparison is shown as in the Table 4. The precision rate is judged by inspecting whether the correct answers, which were supplied in the QAC-2 task, exist between the top 200 of the answer returned by the system.

Table 4 The comparison of the retrieval performance

Retrieval method	Precision rate of questions for common nouns (%)
Boolean model	22.2
Noun-CBVSM	79.0

The results of this experiment shows that the retrieval using the proposed method generates better precision rate comparing to the Boolean retrieval method.

As in the Boolean model, a particular document will only be considered to be related to the question sentences if every single word contained in the query words exists in that document. For many documents, even though they are relatively synonym to the query topics, instead of using the words consisted in the query words, synonym words are used. This phenomenon caused the document not to be considered as related to the query topic. Concerning these topics, from a different point of view, the selection of the query words from the user's query sentence is very important in Boolean model. However, in reality it is quiet difficult for a system to discern the word that is should or should not be chosen as the query word. Comparing to these, the concept-based model enables the retrieval of the significant document although the query words is not included in the document.

V. THE ANSWER WORD EXTRACTION USING THE CBVSM

Even though the candidate documents are retrieved successfully in the retrieval of the related documents, the correct answer will only be extracted if the method of extracting the answer words works properly. In many of the present QA system, the extraction method is created by using a rule-based method. One of the most commonly used methods is named entity extraction method, such as the Named Entity Extraction Tools (NExT) [6]. This type of software works by defining the proper-nouns consisted within the candidate documents. The answer words will then be extracted by referring to the type of answer that is expected, which is formerly defined in the question separator module.

The method of using the NExT is results in excellent performance for the questions demanding proper nouns as the answer word. Unfortunately, as it appears in its name, this method is created to work on proper nouns but not for common-nouns as the system cannot define the common

nouns consisted in the documents. This makes the present system relatively inappropriate to answer questions demanding common nouns as answers.

Secondly, the implementation of the rule-based method to extract the answer, almost depend on the ability of the question classifier to define the type of word that is needed. Moreover, most of the question classifier module is created manually by defining grammatical rules of the sentences. This method works superbly if every single grammatical rule can be defined completely, but it will not be an ideal method as it takes times and costs.

Here, the method of extraction by CBVSM is proposed as one of the solution.

In the concept base, the words and the documents (sentences) are represented by vectors, which the relations between the each words or sentences are calculated. This means that if we have a single word, we can compute out the most related word or the most related sentences to the particular word. Using the same theory we also are able to compute the most related word or sentence to a particular sentence (refer figure3).

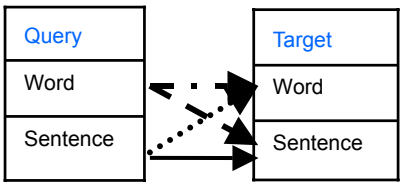

Figure3 The retrieval of the CBVSM

The answer word extraction proposed in this system is based on the theory above. The similarity of the query sentence and the word content in the related documents is calculated to find the word that is most related to the query sentence. The process is done by calculating the similarity degree between the query words and the words contained in the concept base. The word that has the highest similarity degree points its relation to the query topic and has most probability of being the correct answer to the question.

The evaluation of the extraction performance is quiet difficult for the QA system since the performance of the extraction of the answer words depends on the precision of the related document retrieval section. For that reason, the evaluation of the answer word performance is executed by evaluating the entire system.

VI. The Performance Evaluation of the System

The result of the proposed method is evaluated by calculating the precision rate for the 5 answer candidates outputted by the system. As a comparison purpose of this experiment, the result of using the straight CBVSM to retrieve the related documents using the NExT to retrieve the answer word and the result of SAIQA [7] system in the QAC-1 task is implemented (SAIQA is a NTT's QA system that relatively perform a good results in the NTCIR QAC-1, a workshop on evaluating QA system performance).

Table 5 the performance evaluation of the system

	Precision rate [%]	
Extraction Method	Common nouns	Proper nouns
Proposed method	**33.3**	**17.5**
conventional CBVSM + NExT	5.1	11.3
SAIQA	6.1	54.3

This result shows the proposed method generating higher precision rate comparing to the other extraction method for the questions demanding for common nouns as the answer.

XIV. CONCLUSION

In this paper, the QA system with highest precision of the tree tested system for answering the question demanding for the common nouns as the answer word was created. This system was designed by using the Concept-Based Vector Space Model (CBVSM) in 2 sections. The 1st section is for the retrieval of the related documents against the natural language formed question. In order to ensure the performance of the retrieval system, 3 improvements ideas was proposed; the usage of the Sentence-Unit Concept-base, the re-generation of the compounds words and the generation of the Noun-CBVSM. In the second section, the CBVSM was proposed to be used to extract the answer words from the candidate document. As a result, the system succeeds to achieve about 33.3 % of precision against the question demanding for the common nouns as the answer word.

References

[1] T.Kato, S.Shimada, M.Kumamoto, and K.Matsuzawa: Idea-deriving information retrieval system. Proc. Of first NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, 187-193, 1999.

[2] T. Yukawa: An Expert Recommendation System using Concept-based Relevance Discernment. Proc. Of ICTAI 2001,,2001

[3]Y.Matsumoto, A.Kitauchi, T.Yamashita, Y.Hirano, H.Matsuda, K.Takaoka, M.Asahara: Japanese Morphological Analysis System ChaSen version 2.2.1,, 2000

[4] S. Deerwester, S.T. Dumais, G.W. Furnas,et al. Indexing by Latent semantic analysis.Journal of American Society for Information Science, 41(6):391-407, 1990

[5] H.Schuetze, J. O.Pederson : A co-occurrence –Based Thesaurus and Two Applications to Informations Retrieval, ISTL Technical Report ,No.ISTL-QCA-1994-03-02,1994

[6] F.Masui, N.Suzuki and J.Fukumoto: Development of the

Name Entity Extraction Tools (NExT) for text processing. The 8-th Language Processing Society's annual journal, 176-179,2002.
[7] Y.Sasaki, SAIQA-□:A trainable Japanese QA system with the SVM,Journal of IPSJ2004, 2004
[8] G. Salton and C. Burkley: Answering bulletin board questions with referrals. Proc AAAI-96, 10-15,1996