Question Answering System Using Concept-Based Vector Space Model

Isrami Ismail Takashi Yukawa
Nagaoka University of Technology, 1603-1, Kamitomioka-cho,
Nagaoka-shi, Niigata 940-2188, Japan
isrami@stn.nagaokaut.ac.jp yukawa@vos.nagaokaut.a.jp

Abstract

This paper presents the architecture of the Concept-based Vector Space Model Question Answering System (CBVSM-QAS) developed at the Nagaoka University of Technology (NUT) and used in the 4-th NTCIR workshop Question Answering Challenge (QAC) evaluation. The CBVSM-QAS runs on the factual question, which corresponds to the subtask-1 in the NTCIR-4's QAC. One major concept of this system is the idea of placing the whole data set in the concept -based vector space, and searching of the answer for each question is done by calculating the nearest newspaper's document vector. In this paper, the architecture of the system used during the formal run of NTCIR-4's QAC and the architecture of the system after some improvement is introduced.

Keywords: *Concept -based, Vector space,NTCIR*

1. Introduction

Our team, NUT, based in Nagaoka University of Technology's knowledge system laboratory, challenges to study on the concept of concept-based model and its potential in information retrieval. The reason why we focused on this concept-based model is that the model is theoretically able to retrieve information despite the incompleteness of the query words. As for the presently used information retrieval algorithm such as Boolean model, only the document which contains all of the query words will be retrieved (in case of the query expression is composed using AND operator). This means that wit hi n all of the query words, even if one word is not contained in the document; the document will be removed as a candidate of retrieval. This will probably result in some important documents being dropped out during the information retrieval. Concerning this problem, we are attracted by the potential of concept-based model. In the concept-based model, information retrieval is executed by comparing the query characteristic to document characteristic that is even some words in the query words are not contained in a particular document. The document is still judged as one of the candidate answer.

In this paper, implementation of the concept-based vector space as the main algorithm for the question answering system is proposed.

2. The generation of concept-based vector space.

The concept-based information retrieval model is one of the variations of the vector space model. To put it simply, the concept-based model is a knowledge-based of the words contained in the document set, and vectors are assigned to each word. In the concept -based vector space model , each target document is characterized by the appearance frequency of each word contained in a particular document . This means that the concept –based contents differs for different document collections. Document vector is generated by the composition of the words' vector contained in the document. Using this document vector and the vector of the question word, the cosine' s coefficient is calculated and ranks the entire target document. In this IR model, the documents with the larger cosine coefficient should be the documents which are most related to the question word. The document poses the nearest location to the question word in the vector space comparing to other documents. This explains the merit of the concept -based model as even some words in the query vector are not contained in the particular document, as long as the cosine coefficient is the biggest, the document is the best candidate answer.

In the concept -based model, as the word vectors are generated using the target documents, the words' co-occurrences in the target document s are statistically calculated. In detail, for N words (N=1,000~10,000), which is the word with high frequency contained in the target documents, a neighborly co-occurrence matrix (N×N) of the words is generated. (Refer to table1). Due to the limit of the amount of computing resources, the words contained in the concept-based are limited to 10,000 words. The words are chosen with the priority on the words with high co-occurrences frequency. The words with lower co-occurrences frequency in the entire document collection will have a higher probability of dropped out from the concept-base.

	Word A	Word B	Word C	…
Word A	0	0	1	
Word B	0	0	0	
Word C	2	0	0	
:				

Table1: generation of the N×N Matrix

This matrix then, compressed to 100~200 dimension by implementation of Singular Value Decomposition (SVD), finally results in generating the concept base (Refer to table2).

	c_1	c_2	c_3	…
Word A	c_{11}	c_{21}	c_{31}	
Word B	c_{12}	c_{22}	c_{32}	
Word C	c_{13}	c_{23}	c_{33}	
⋮				

Table2: compression of N×N Matrix by SVD (Concept base)

Document vector, which represents the concept (words) included in the particular document, is generated by the composition of the concept vectors within it. The question vector is similarly generated by the composition of each concept contained in the questions, a short sentence composed from some words.

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

The query sentence supplied by the user is also a set of words. Therefore, the query can also be considered as one type of document. Query vector, with the same concept of the document vector is represented by the words vector composed in a particular query sentence.

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

In this case, the query vector will be generated every single time when the user mentions a new query to the system. Before the generation of the query vector, the stopwords will be eliminated in advance, identical to the method of the document vector generation. This process put both of the document vector and the query vector under a totally same conditions so that the comparison of these entire vector become possible.

The Retrieval process is done by calculating the similarity degree of the question sentence vectors and the each document vectors. The target document are then sorted according to the degree of their similarity against the question sentence for the most related document should be the document with highest similarity degree.

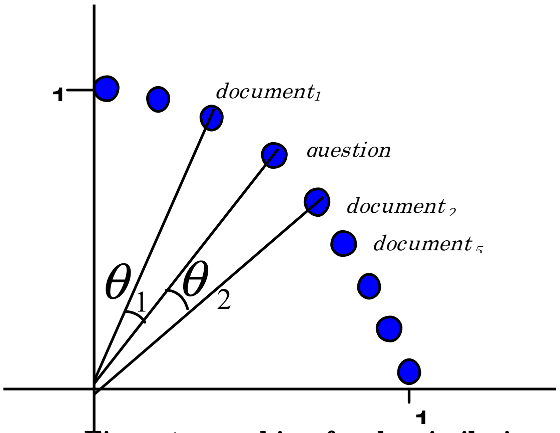

Figure1: searching for the similarity degree of the question and the document collections.

3. The architecture of the CBVSM-QAS

The NTCIR-4's QAC was the first attempt for NUT team's to design the question and answering system. By that reason, many defects were created within the progress. NUT's attempt to implement the CBVSM into the question answering system continues but by the reason of time limit of the NTCIR-4's QAC2 formal run, we will explain the architecture of the CBVSM-QAS on the first stage (that used for the QAC2 formal run), and the improved system (created after the QAC2 formal run deadline).

3.1 The first generation of CBVSM-QAS

The first generation of the CBVSM-QAS was created using the CBVSM model to extract the candidate documents from the entire newspaper database collection (Mainichi Newspaper collection year 1998 & 1999 and Yomiuri Newspaper collection year 1998 & 1999). The retrieval for the answer word is done by the Name Entity Extraction Tools (NeXT, developed by Masui,Suzuki and Fukumoto[1]). The architecture of the first CBVSM-QAS is illustrated in Figure 1. The architecture o f the system can be separated into 4 main sections. Section A is the section of generating the concept-based vector. Section B is the section of processing the question. Searching process for candidate documents is done in section C, while the extraction of the answer word is done in section D of the system.

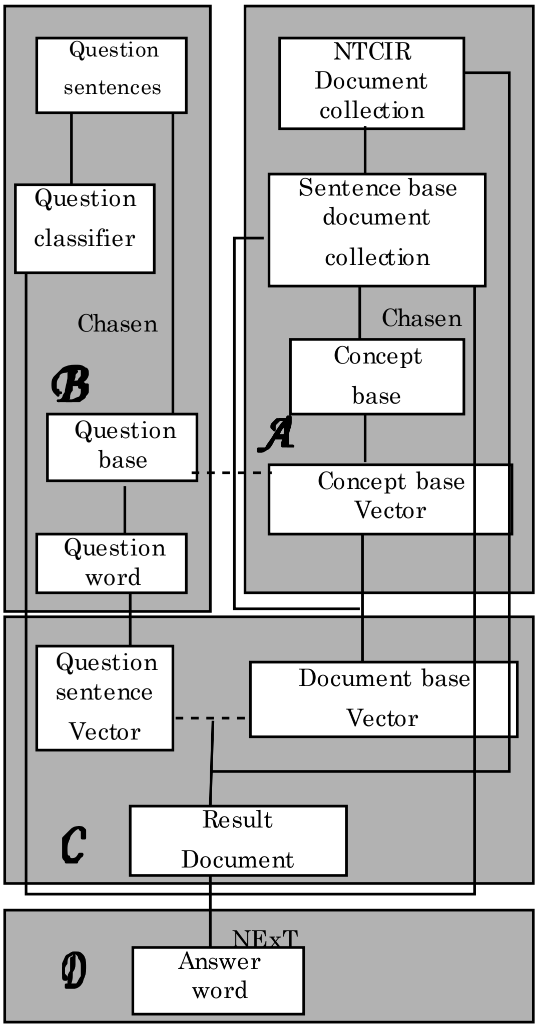

Figure1: 1st generation of the CBVSM-QAS

It was found in our study that the accuracy of the retrieval almost failed if the calculation of the document vectors was done directly for each existing document. In the raw newspaper documents, the document size is too large that the words characteristic (word vectors) for the words contained in the question sentences are not emphasized in the document vector. In order to solve this problem, the sentence-based document method was proposed (refer section A of Figure 1). Each document was divided to a sub document composed of each sentence from the raw documents. The sub-documents then go through the morpheme analyzer of Chasen[2] and a concept-based is created. Before the vector of each word (concept-based vector) and the vector for each sentences base document s (document base vector) are calculated using the method explained before.

The question sentences, as given in the NTCIR's QAC2 subtask 1, will then also be processed using the same method, in order to produce the vector for the question sentences. After the question sentences are processed by the Chasen to generate the question vector, the words contained in the question vector are compared to the words in the concept -based. The words that are contained in both will be selected as the words to be searched (question words). The words contained in the concept-based are limited to 10,000 words due to the limit of the amount of computing resources. This means words that are not contained in the 10,000 words of concept-based will not be considered as question words. At the same time, the questions are processes through the question classifier to identify the question type(section B). The question types are divided as below;

- Person: the question appointing to the person's name
- Organization: the question appointing to the organization / association name.
- Location: the question appointing to place/ location
- Time: the question appointing to specific time or a period of time.
- Money: the question appointing to answer concerns with the currency of some countries.
- Date: the question appointing to day or date.
- Accuracy: the question appointing to percent

ages of some numeral data.

- Others: the question appointing subjects other than classified above.

As mentioned in chapter 2 of this paper, the searching process for the answer document is executed by comparing the question vector and each document vector (section C). The system will then sort out the similarity degree and select the top 3 of the results, which is referring to the sub documents which are most related to the given questions.

The result documents, after extracted from the main document collection are then processed through the NExT. In this process the proper noun contained in the text will be extracted (section D). Referring to the result of the question classifier, the answer word will be chosen from the extracted nouns.

3.1.1 The performance evaluation of the first generation system

Because of time limitation, the first generation of the CBVSM-QAS worked with very limited performance. The main cause of the poor performance of this system was the shallow knowledge in creating the rules of question classifier and the system's failure to recognize the answer word from the extracted documents. On the question classifier, about 35% of the questions are hardly classified because of the weakness of the classifying rules implemented in the system. For the questions that succeed to be classified by the question classifier, the system will depend on the performance of NExT to identify the answer word. As the NExT only works with the proper nouns to be identified, common nouns will probably be abandoned by the system. Poorly, the system only achieved 0.018 of system score on the NTCIR-4's QAC2 subtask 1. However, the attempt of creating a question answering system with the implementation of concept -based vector space model by NUT continues, as we believe good answers can be achieved if only the method of question classification and extraction of the answer word could be improved. As a strong support for our belief, the concern document (the document that contains answer for the questions) in fact, did come to the top of the list during the comparison of the question sentences' vector and the document base vector performed during the searching process.

3.2 The second generation of the CBVSM-QAS

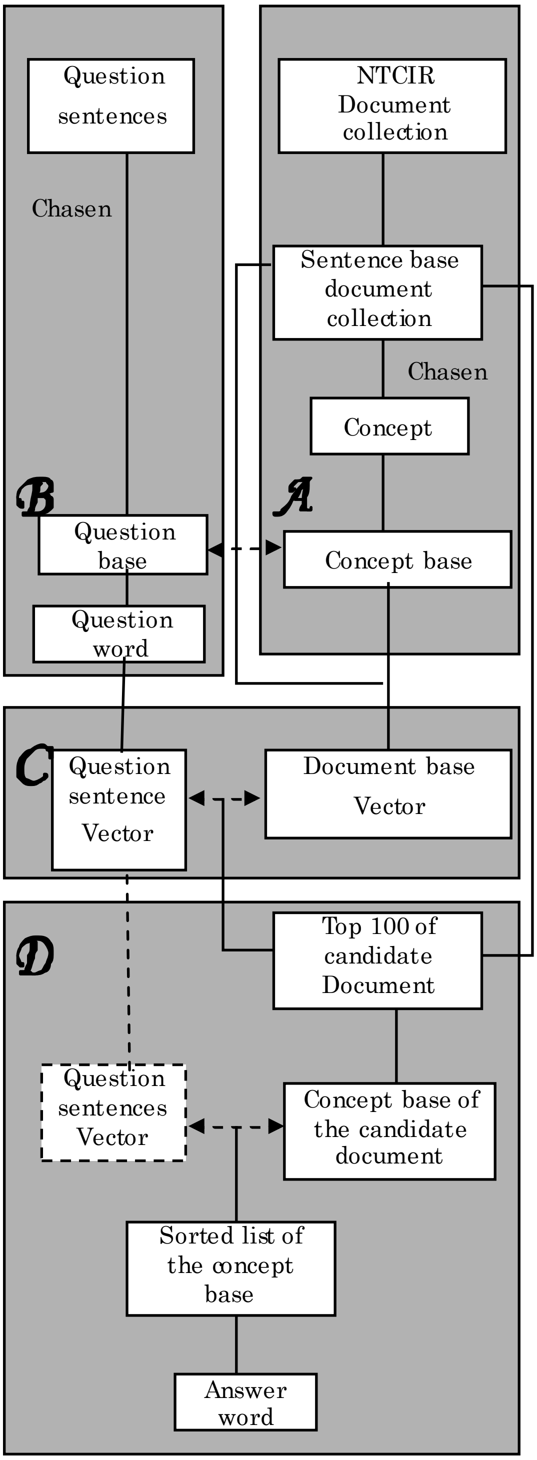
Question sentences
NTCIR Document collection
Chasen
Sentence base document collection
Chasen
Concept
B
A
Question base
Concept base
Question word
C
Question sentence Vector
Document base Vector
D
Top 100 of candidate Document
Question sentences Vector
Concept base of the candidate document
Sorted list of the concept base
Answer word

Figure 2: 2nd generation of the CBVSM-QAS

Realizing the failure of the first generation system, the second generation system was built with some modifications on the system's structures. The question classifier in which is hard to fix the classifying rules has been pulled out from the second system. The search process still uses the same concept as previous one. Since the first CBVSM-QAS was dependant on the NExT performances, the second generation system was improved by the implementation of a new method of extracting the answer word. The system's architecture is explained by Figure 2.

Section A of this system, which refers to the generation of concept-based and document base vector remains the same as the previous one.

Since pattern matching was used to classify in the first generation system, a very complicated pattern matching rules is needed to cover any type of question. These conventional ways of setting hundreds of rules of pattern matching would surely work properly only if the rules can be written perfectly, but it needs a great effort to complete it. Instead of using the question classifier, we have modified the system by skipping the question classifier and implementing the 100-documents concept-based retrieval model during the answer word extraction process. The 100-document concept-based retrieval model was also expected to gain better results compared to the former structure. Accordingly, the question sentences, without being processed through the question classifier are directly converted to question sentences vector by the same method as in the first system.

During the search process (section C), after the comparison of the question sentence vectors and the document base vector is executed, the system will choose out top 100 of the related documents instead of choosing the most related document as in the answer document as in the previous system.

The selected documents, which are referring to the sentence base document, then will go through the process of generating the concept-based of its own collection. Since this collection differs greatly from the main collection in the concept of words contained, almost of the words contained in this document collection are able to be stored in this concept-based without bothering the limit of 10,000 words of the concept-based. Here, the question once again will be checked over the new concept-based to choose the words that are included in the new concept-based as the question word. This is conducted due to the differences of the words contained in the whole document concept-based and the 100 document concept-based. Some of the words especially the words with low frequency of appearance will probably be dropped out from the whole document's concept-based are able to take part in the 100-document concept-based. Accordingly, a more accurate value of the question sentences vector generated. The question sentences vector and the concept-based vectors of the 100 documents then will be used to calculate the scalar product of the question sentences and each word from the concept-based. The main idea is base on the word which is having a close value (bigger cosine co-efficient) to the question sentences might be the answer word of the question.

3. 2. 1 The performance evaluation of the second generation system

In this second generation system, the question given by the user was not classified for the answer searching purpose as we have realized that the classification of a question need to be really perfect as will effect the system performance.

The second generation system was able to process the question querying for dates. For the question such in the QAC2-10003-01, "When was Shochiku, a long-established group in the theatrical world, founded?" the system succeed to extract the answer as " year 1902". Comparing this to the first generation of the system, the first generation was totally out because the system was unable to classify the question.

For the question that concerning person's name as in the question QAC2-10009-01, "What is Onodera Shotaro's real name?" proper answer still cannot be extracted by the system. The main reason was that for this question, the words that were selected as query words (in fact, the words that is contained in the concept -based of the entire documents) were only the words "what" and "real name". It seems that the word "Onodera Shotaro" appears in the entire newspaper document only several times and the appearance frequency of the word is too small to be selected in the concept-based. This is results the unrelated candidate document was extracted from the database. The words that containing the words "what" and "real name" were selected without considering whether the word " Onodera Shotaro" is contained in the particular document or not. The problem also appears in the question QAC2-10010-01, "Where was Ishinomori Shotaro born?" which only the words "where" and "born" was selected as query words and question QAC2-10012-01, "where does the writer, C.W.Nicole, lives?" which only the words "where" and "writer" was selected.

For questions concerning the place and organization, the system shows different performance for different question. For some of the questions, the system somehow was able to get the expected answer to top 3 of the answer word (referring to the NTCIR-4 QAC2's subtask 1, the candidate answers were submitted as 3 candidate answers. For quest ion QAC2-10103-01," The highest peak in the seven continents of the world is Mr. Everest. Which mountain is the highest peak in each continent?" the answer "Kilimanjaro" and "McKinley" returned as one of the top answers.

For the questions classed as "others" in the first system, the second system also shows inconsistent results as for the question QAC2-10124-01, "What are the names of the satellites of Jupiter?", the systems returns the answer " Europe" as the first candidate. Unfortunately, the system failed to answer other questions such as question QAC2-10183-01, "What are Misora Hibari's famous hit songs?".

4. Conclusion and future plans

From the result of the attempt of implementing the CBVSM into the question answering machine, the characteristic of the CBVSM had been confirmed. The CBVSM-QAS performs well if most of the nouns contained in the question sentences are selected as query words, which means that all the query words are contained in the concept-based. The CBVSM-QAS will perform poorly in the reverse situations.

In the query words-fully selected situation, inaccurate vector might be produced due to the compound words problem.

Our research continues on how to improve the performance of the CBVSM-QAS. Some of the plans are already proposed and going to be implemented and evaluated in the CBVSM-QAS.

1) Generation of noun-based concept-based model

In the question answering system, whether in the question sentences nor the in the answer word, the most useful words are nouns. Nouns should be the most important words in the question sentences. Nouns also are the state of most of the answer words. This means that by eliminating the words other than nouns, the similarity degree of the question vector and candidate document can be improved. The elimination also will somehow solve the limited number of words in concept-base by placing only the important words in the concept-base. The elimination of the words is executed by sorting the words into verbs, nouns, post positional practical words and adverbs. Only the nouns should be selected to be inserted into the concept-base. By the implementation of the noun-base concept-base the number of nouns contained in the concept-base certainly will be increased.

2) The revision of the compound words' vector

Phrases such as "shudanboukoujiken" (group violence incident), or the words such as "kaseit ansaki" (Mars investigation plane) are presently divided into separate words vectors. "group violence incident" is a phrase that refers to a specific incident that should be taken as one word. Unfortunately the phrase is present l y separated into "shudan"(groups), "boukou"(violence) and "jiken"(incident). These kinds of words are massively constructed in the concept-based especially for the phrase that refers to technical terms.

In the concept -based, words' characteristics are represented by the neighborly co-occurrences between the words. For that reason, if the morpheme analysis of the words is not performed properly, the neighborly concurrences between the words cannot be represented correctly. This might cause some inaccurate vectors to be retrieved for the particular query. The problem should be able to be solved by identifying the compound words before the construction of the concept-based.

5. References

[1] 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, pp.176—179,2002.

[2] Y.Matsumoto, A.Kitauchi, T.Yamashita, Y.Hirano, H.Matsuda, K.Takaoka, M.Asahara

[3] T.Kato, S.Shimada, M.Kumamoto, and K.Matsuzawa, Idea-deriving information ret rieval system. Proc of first NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp. 187-193, 1999.

[4] R. Baeza-Yat es, B. Ri bei ro-Net o. Modern Information Retrieval, ADDISON WESLEY, 1999.

[5]S. E. Robert son and K, Sparck Jones, . Relevance weighting of search terms. Journal of the American Society for Information Sciences, 27(3):129-146, 1976.

[6] Boris Galitsky, Natural Language Question Answering System, 2003.