Pak Pak Serves You
<<Application on Interactive Dialogue Question Answering>>

PUN Ka I
Faculty of Science and Technology of University of Macao (Macao S.A.R)
punkai91@yahoo.com.hk

LEONG Pok Man
Faculty of Science and Technology of University of Macao (Macao S.A.R)
luisleong@gmail.com

MAK Hoi Fong
Faculty of Science and Technology of University of Macao (Macao S.A.R)
leoagneau@yahoo.com.hk

WONG Fai
Faculty of Science and Technology of University of Macao (Macao S.A.R)
derekfw@umac.mo

LI Yi Ping
Faculty of Science and Technology of University of Macao (Macao S.A.R)
ypli@umac.mo

Abstract

Question Answering is a type of information retrieval. Then, Speech Question Answering is a kind of information retrieval which requires complex natural language processing techniques together with the speech recognition techniques. Most probably, QA system is used at web site by giving certain questions as the input and the web server will try to retrieve the answer from the database. Besides, there is another type of informational retrieve system called telephony IVR (Interactive Voice Recognition) system which is very commonly used. Most of the question answering system which is probably used is somehow a unidirectional communication with the user, that is, the system is not able enough to have interaction with the users, however, there is a trend of developing a bidirectional question answering system with some interactive dialogues as the importance of artificial intelligence is running up. Here, our research is following this trend by using speech. Here, our focus will be on researching this field and let the East Asian Games (EAG) be the domain of our research. That is, the spoken question will be recognized by our speech engine and our system will extract a correct answer, or gives some other communicative dialogues if the question is out of our domain, in response to the questions spoken by the user. There are two main parts in our research: one is QA system and the other one is the speech recognition development.

Introduction

This paper describes a study which is on the field of QA system using the technique of speech recognition. In the part of QA system, we put our focus on the analysis of the structure and the components of the implementation of the system, and also what the principle is as well as what and how the techniques will be used in it. For the part of development of speech recognition, we will analyze the speech engine and discuss some of the difficulties and restrictions we have met, for example, the ambiguity like uttering, different words but with same pronunciation, low accuracy when dealing with large amount of words and so on.

QA system can be divided into two specific parts of closed-domain and open-domain[1]. The difference between them is that the previous deals with questions under specific domain question answering and therefore for the knowledge about the questions can also be specific and whereas for the open-domain question answering deals with almost all kinds of questions and therefore the amount of data from which the answer extracted will be much larger. In our research, we have chosen to use the closed-domain specific question answering system and our focus is in the field of East Asian Games (EAG) which is held in Macau in 2005. By using this, we can increase the accuracy of the speech recognition and decrease the ambiguity since the amount of the words in the database used is greatly reduced.

The paper is outlined as the following: the next section is focused on the analysis of the QA system Architecture with speech recognition engine. After this, we will discuss how we make use of the IBM speech engine to work with our system. Then, we will describe how we analyze the question, how to deal with the matching and construction of question and answer. After all, we will talk about the difficulties met and also our future work of this system. Finally, we will give the conclusion of the paper.

Interactive Dialogues QA System

For us, the human being, we usually actively confirm with the user that the question he/she has entered is correct or not, or we may try to get some keywords from the question he/she given in order to organize another question which is somehow related to the question input by the user and try confirm with the user with that question. Such kind of communication of course required many dialogues with the user, because of this, there are several problems in developing the system.

Issues that to be addressed in developing IDQA System

Before introducing the architecture of the IDAQ System, we must first identify the problems which we must meet during the development of the system.

Speech Recognition Accuracy – The accuracy of speech recognition is not quite satisfied. Without filtering, the accuracy can be only 20% empirically. Although we have design a set of grammar to deal with the problem of low accuracy, this problem can only be lightened but could not be solved completely. There are actually lots of programming work to deal with this problem, like if the question could not be fully recognized, the system must be able to interact with the user. The system should be able to interact with the user in order to get more information so as to retrieve the answer for the user.

Human speech – Human has numerous ways to ask the same question with the same meaning. The system should understand most possible ways in asking a question. For example, people have at least 10 ways in asking “How to go somewhere”. For a polite person, many courteous words may inside the question, for example, “please”, “I would like to…”, “thank you”, ‘Excuse me”. In human communications, those words are needed and are approved; however, for machine-based system, those words are somehow obstacles. Speech QA System needs to accept and filter out those words in order to minimize the effect in recognition process.

By solving the above problems, unambiguous question can be got by the system and it can follows an interactive approach, which is discussed below, to achieve the goal of interacting with dialogues with the user in the closed-domain QA System

Architecture of the system

After referring to the earliest QA System Algorithm presented by Simmons at 1973[2] and many of the publications in TREC QA System[3], our QA System has been divided into five modules as shown in figure 1:

i) Speech Controlling Module

ii) Speech Analysis and Matching Module

iii) Answer Construction Module

iv) Speech Synthesizing Module

v) Knowledgebase Module

At the beginning of our paper, we have mentioned that the QA System will be developed with the techniques of the speech recognition in order to let the user can input their question with their speech. Here, the first module has taken the role of getting the question input by the user with the Speech Controlling Module. The recognition engine of this module is IBM Speech Engine due to their 20-years profound research and development in that area. Besides, IBM has provided a complete API for developer so that we do not need to spend so much time on the low-level controls of the speech engine and recognition processes. After getting the input question, we need to pass that question to the Speech Analysis and Matching Module to analyze the question. If it is successfully analyzed, it will be passed to the Answer Construction Module to construct the answer; if not, or corresponding answer can not be extracted, response will be given by the system through the speech engine and therefore communicate with the user so as to achieve the goal of bidirectional communication between the user and the IDQA system. Answer will be extracted from the Knowledgebase Module, just like database, and then pass the data to the Answer Construction Module and Speech Synthesizing Module to construct the answer for output.

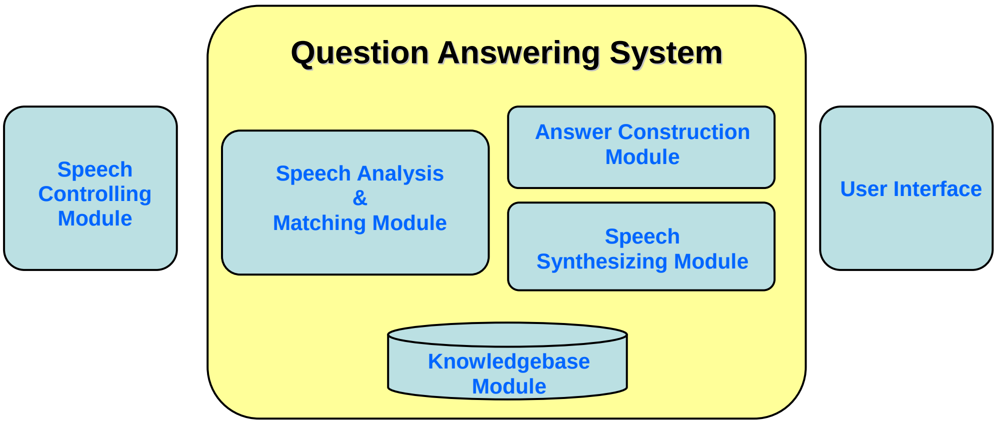

Question Answering System
Speech Controlling Module
Speech Analysis & Matching Module
Answer Construction Module
Speech Synthesizing Module
User Interface
Knowledgebase Module

Figure 1 – Architecture of Interactive Dialogues Question Answering System

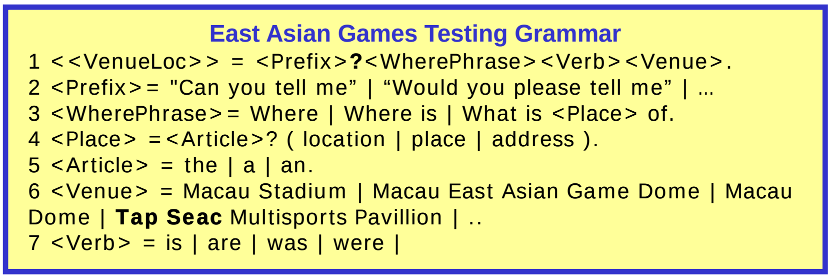
Questions filtering by IBM Speech Engine

Nowadays, the overall accuracy of speech recognition is not good enough to recognize all the speech from all different kinds of people. Some of the speech engines, for example, IBM, Microsoft, need a process called “Training” in order to increase the accuracy. If this QA system is single-user oriented, the training process can enhance the accuracy of the speech recognized. By using this process, user needs training the speech recognition engine by reading some passages so as to let it get familiar with different ways of speech of the user. However, as we can see, this training process will just improve the accuracy of that user, and this kind of speech recognition is called **user dependent** recognition.   For IDQA system, which is a system for multiple users, this training process may be meaningless and unreasonable since we could not ask the users to train the engine in advance. This kind of recognition is called **user independent** recognition. For this reason, we decide to use the IBM Speech Engine without training.

In order to know the rate of accuracy of the IBM Speech Engine, we have done some testing and we found that the accuracy without training can be very bad. IBM Speech runtime can recognize more than 200,000 words including professional and special nouns. For short sentences, the accuracy is acceptable. For longer sentences, or sentences include proper nouns or **homonyms** words(feel, fill, field), the result is poor, in some cases, the recognized words can be even **unpredictable**, like, “Macao” is recognized as “Moscow” very often.

Due to the reasons we have discussed above, some restrictions about what can be said must be raised, i.e. users can only ask the questions within certain field (domain of knowledgebase). For example, if the domain is inside the campus, the proper nouns can be “library”, “lab”, “plaza”, “toilet” and many other else, other irrelevant proper nouns or homonyms are omitted.

We can achieve the method we have discussed above by designing the grammar provided by the IBM speech engine, the Speech Recognition Control Language[4]by ourselves. SRCL Grammars formally define the set of allowable phrases that can be recognized by the speech engine, the language is in EBNF, such as:



East Asian Games Testing Grammar

```
1 <<VenueLoc>> = <Prefix>?<WherePhrase><Verb><Venue>.
2 <Prefix>= "Can you tell me” | “Would you please tell me” | ...
3 <WherePhrase>= Where | Where is | What is <Place> of.
4 <Place> =<Article>? ( location | place | address ).
5 <Article> = the | a | an.
6 <Venue> = Macau Stadium | Macau East Asian Game Dome | Macau
Dome | Tap Seac Multisports Pavillion | ..
7 <Verb> = is | are | was | were |
```

Figure 2 –Sample SRCL Grammar

In figure 3, “|” means alternation and “?” has the meaning “with or without”.   For example, in recognizing the question “Where is Macao Stadium”, this question is in our specified grammar. This simple grammar is an example; the real grammar is much more complex.

These phrases can be spoken continuously, without pausing between words. The grammar provides a language model for the speech engine, constraining the valid set of words to be considered, and increasing recognition accuracy while minimizing computational requirements. In our Q/A System, this is a possible way to filter out unpredictable sentences. Figure 2 shows how the speech engine works with the other modules in the system.

After applying the grammar, we can notice that the recognized outputs now contain no unpredictable words anymore, i.e. all words are under expectation. Synonyms problem has also been lightened because I just defined “hi”, “list” in the grammar but not “high” and “least”. Grammar simplified the implementation of the system because we do not need to do many associations on synonyms and unpredictable words.

It does not mean the problem on unpredictable words and synonyms has been solved, some words may also be recognized incorrectly, for example, “When” may still recognized as “Where” or “Where is” or even “Wednesday”. Fortunately, the lightened problem can be even lightened in Speech Analysis Module by associating during matching phrase.

Analysis Routine in QA System

In order to make the real-world users find the QA System is useful, several standards must be met[5]. They are timelines, that is, the answer to a question must be provided in real-time; accuracy is extremely important because wrong answer is even worse than no answer; moreover, the answer is usable for the enquirer or not for there may be some cases that different answers are needed for that same questions for different domains; whether the complete answer is provided to the enquirer or not; and also the answer provided must be relevant within a specific context. Therefore, in order to fulfill these standards, a correct analysis of the question must be given before all. By referring to some passages and papers[6][7], we have concluded the following algorithm in analysis the problem.

i) coarse analyze the input question to identify its type (types include time, location, who, amount, method, size, number, what, command and determination) and simplify the question; if the system is unable to get the question type, communication with the user again so as to get some keyword

ii) identification of keywords and generation of the input question pattern after the simplification of the input speech

iii) matching of the input question pattern with those predefined in the database will be done in order to select a champion one to construct a corresponding answer

iv) detailed analysis according to selected question pattern from the previous step to identify the answer format, which is predefined, provided to the user

v) answer will be generated according to the data from the database and the format identified previously

Here, we first give the analysis procedure of dealing with normal question

After the recognized speech from the Speech Controlling Module is processed, the correct input speech from user can be extracted. Moreover, by currently stage of the user, corresponding speech index is assigned to that input speech. There are two kinds of information: the extracted input speech and the speech index. They will be passed to the next module, that is, Answer Construction and Speech Synthesizing Module, to generate the output from the Commands and Greetings; whereas for the input speech recognized at the QA Stage, that is, Question and Confirmation Speech will be passed to the next section, Speech Analysis, for further processing.

Meaningless words are filtered lighten the processing workload of the system in the Speech Analysis and Matching Module. Other ambiguous words and aliases of the keywords have also already be linked together to enhance efficiency and performance. Furthermore, with the keywords and their types, the input speech pattern has been generated and passed to the next section for matching the standard speech pattern defined in IDQA System. However, in case of the Speech Type of confirmation answer, some additional information like confirmation type and confirmation flag are also needed to pass to next section.

The matching rate of the input speech pattern for each of the defined speech pattern is calculated according to their similarity. Then, the speech pattern with the highest speech will be selected to the construction of different types of response according the level that the matching rate can reach. Finally, the output of this section, or this module, to the next module will be the input speech pattern and also a flag that indicate which response should be generated.

Coarse and ambiguous utterance Analysis

The analysis of ambiguous utterance for computer can be very complex. Since the accuracy of speech input is not that high, wrong input may be generated very often. Empirically, human often think and pause during asking questions, then some irrelevant words what we have just mentioned will appear. So, the system should filtrate out those words and make sure it got the complete question but not a partial one.

What difficulty about answering questions is that the fact that before a question can be answered, it must be first understood. One level of the interpretation process is the classification of questions. This classification should be determined by well defined principles. As mentioned, wrong answers are worse than no answer.

If the system does not understand the question entirely, the best answer is “Sorry, please ask again”. How about the system understands the question partially but not very well? We have used the method just like human being, that is, our system will try to repeat the question which has already been modified to the user in order to get the confirmation. Then, just as the procedure above, we can either proceed to the next module or start the procedure all over again.

Answer Construction Module

In order to make the system usable and also provide a user-friendly interface, the quality of the output of answer of user’s question, the introduction of the system, plus the user guide to user is important to the usability of the system. This module is to construct all the output of the system. As mentioned above, one of the characteristics of the system is to be interactive, so providing useful, easy to understand and informative output is very important to user. Giving clear instructions to user can also guide him/her to use the system in a more convenient way, which will benefit to both users’ feeling and to system’s operation.

The main methodology of the construction of the answer is to build the sentence according to “answer pattern”, which acts as the skeleton, or the template of the sentence. The content of the sentence are retrieved from the keywords database, according to the result of processing the input from the output of the Speech Analysis and Matching Module.

The greeting of the system and other introductions and guidance of the system are also generated by this module, which is also according to patterns and keywords defined in the database.

Speech Synthesizing Module

This module uses the Text-To-Speech (TTS) engine provided by IBM. All the audio output which finally be heard by user are generated by this module. The text input of this module will be passed to the engine, after some enhancement of the sentence to make the acoustic effect to be more acceptable, the synthesized text will then be output through the speaker to user.

Difficulties

After getting an unambiguous question, there are still several bottleneck difficulties in the research and implementation of Speech QA system.

Taxonomy – Questions are difficult to be classified; then, if the question cannot be classified, low efficiency and low usability may occur. For a question with no type, it may need to scan through all candidates in the knowledgebase and match against with all of them in order to evaluate the most possible answer. This problem occurs when the question contains “How” or “What”.

Constructing complete answers – Even completeness in QA system is very important, the difficulties in answer complete questions are a bottleneck. Knowledgebase are not presented as passages or documents but in different database tables. Most of the time, answers are dynamically constructed from different tables.

Matching and Scoring the questions – The efficiency and correctness of catching keyword are both low. Setting up an efficient algorithm for algorithm for matching is a very tough job. Till now, our algorithm for catching keywords is not good enough. This is also the most difficult part we need to solve in the research and implementation of IDQA.

We need to solve the bottlenecks discussed above one by one in the future. Particularly, the above bottlenecks are human natures. Human can easily solve them but computer cannot. So, solving the above problems is indeed improving humanity in computer science, an area of Artificial Intelligence.

Prototype of Interactive Dialogues Question Answering System

The main interface of the prototype of IDQA is shown in figure 3. Here we can use our microphone as the input media. Here, PakPak is welcoming us as a guide!

User can chat with PakPak with some greeting words such as “Hello”, “Good Morning”, etc. As shown in Figure (a), a user greet with PakPak by saying “Hello, Good afternoon”.

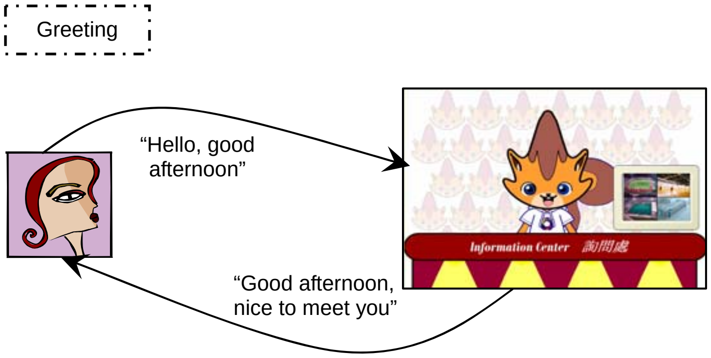


Figure 3 (a) – Main interface of IDQA System (Greeting)

When user wants to ask a question about the 4th EAG, user can ask a question as shown in Figure 3(b)

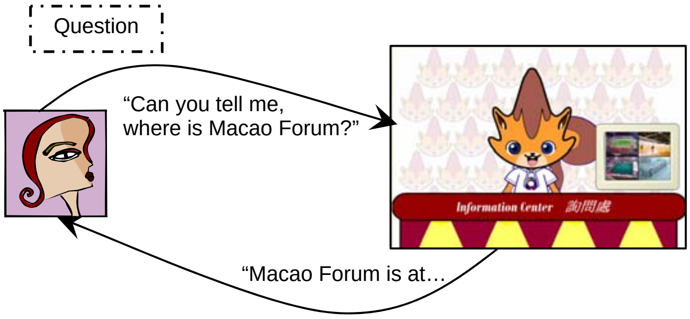

Question
"Can you tell me, where is Macao Forum?"
Information Center 詢問處
"Macao Forum is at…

Figure 3 (b) – Main interface of IDQA System (Question)

If the spoken question is not totally recognized, PakPak will ask back the user to provide him some more information. As shown in Figure 3(c), PakPak can just hear the word "swimming", then he ask back the user what information the user want to know about this competition.

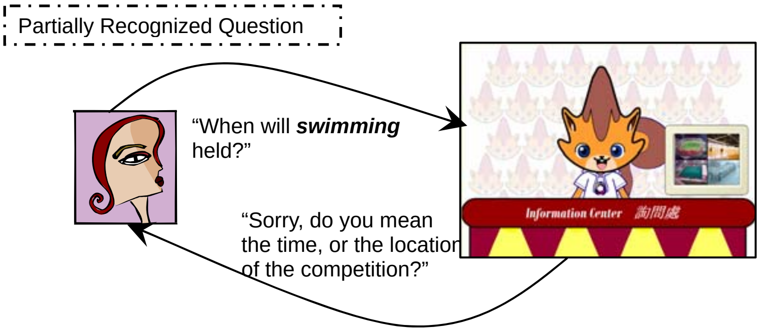

Partially Recognized Question
"When will *swimming* held?"
"Sorry, do you mean the time, or the location of the competition?"
Information Center 詢問處

Figure 3 (c) – Main interface of IDQA System (Question is partially recognized)

In figure 3(d), user replies PakPak by confirming him she wants to know the time of swimming. Then, PakPak replies her by using this information.

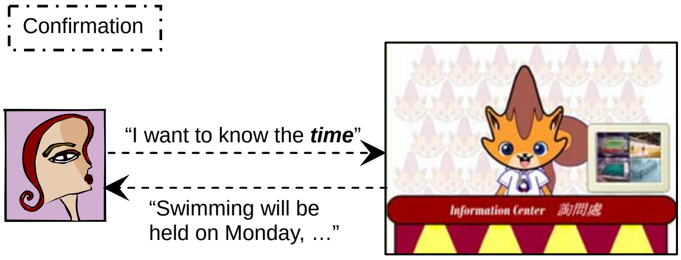

Confirmation
"I want to know the *time*"
"Swimming will be held on Monday, …"
Information Center 詢問處

Figure 3 (d) – Main interface of IDQA System (Question is partially recognized)

Conclusion and Future work

Speech QA System is an interpretation of Human-Computer Interaction. Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them. Human use their speech instead of typing into the keyword and the computer, making computing more natural, like a friend but not a machine. This paper has described a preliminary prototype Speech Question Answering System. At this moment, our system is still not fully-functioned; however, we have already developed some of the basic functions although the functions are not very well-developed. In the coming work, we need to enhance the technique in analyzing the input question; therefore, we need in improve the accuracy in getting the keywords. If the keywords could not be caught correctly, our system would be meaningless to some at extend. Moreover, since the grammar which we are using now is just only good enough to accept the most common ways of question asking, we need to improve the ability of accepting them by enhancing the design of the grammar. Furthermore, an algorithm of matching and scoring the questions must be designed for the seek of getting a correct and unambiguous questions. The algorithm we are using now is only matching keywords in the question and the score is given according to the number of keywords matched. Later on, we are going to improve this by applying Dynamic Programming (DP) with different keyword should has different score. Finally, we are going output the answer with the speech instead of text which we are using so as to achieve the goal of having dialogue with speech.

References

[1] Question Answering, "*Question Answering (QA) is a type of information retrieval*", http://en.wikipedia.org/wiki/Question_answering.

[2] Simmons, R.F., "*Semantic Networks: computation and use for understanding English sentences*", San Francisco, 1973.

[3] Enrique Alfonseca, Marco De Boni, Jose-Luis Jara-Valencia, Suresh Manandhar, *"A prototype Question Answering system using syntactic and semantic information for answer retrieval"*, in Proceedings of the Tenth Text Retrieval Conference (TREC-10), Gaithersburg, US, 2002.

[4] IBM(R) ViaVoice™ SDK - SMAPI Developer's Guide, pages 17-18.

[5] John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, Ralph Weishedel, *"Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)"*, 2001.

[6] Kenneth C. Litkowski, "*Syntactic Clues and Lexical Resources in Question-Answering"*, Ninth Text REtrieval Conference(TREC-9). Gaithersburg, MD. November 13-16, 2000.

[7] Ellen M. Voorhees, *"Overview of the TREC 2003 Question Answering track"*, In TREC 2003 Proceedings, 2004.