Information Acquisition Using Chat Environment for Question Answering

Calkin A.S. Montero and Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University,
Kita-ku Kita 13-jo Nishi 8-chome, 060-8628 Sapporo, Japan
{calkin,araki}@media.eng.hokudai.ac.jp,
WWW home page: http://sig.media.eng.hokudai.ac.jp

Abstract. The main purpose of question answering (QA) is to find an accurate answer to a given question. During research on QA systems in open domain, we understood that, in many cases the information that can be extracted from a question is not enough in order to find a suitable response; to be specific, for the purpose of answering a user's question, additional information is needed to accurately fulfill his/her expectations. In this paper we introduce our idea of information acquisition using Chat Environment for QA and the results of preliminary experiments.

1 Introduction

Human-computer conversation (HCC) is a part of natural language processing technology, one of the oldest, most important, and most current areas of Artificial Intelligence (AI) that has reached a similar stage of development as some better-known areas of language processing, like Information Extraction (IE) and Machine Translation (MT). Another field of AI that has become a powerful paradigm is question answering (QA), extending beyond AI systems to query processing in database systems and many analytical tasks that involve gathering, correlating and analyzing information.

A considerable amount of research has been done regarding to these two fields of AI. One of the most famous examples of HCC is ELIZA [1], a computer program that interviews a psychological patient without limiting words. Another well-known dialogue program is PARRY [2], whose goal is to simulate a paranoid patient. Recently the development of dialogue systems has increased exponentially with advances in areas like dialogue management and context tracking techniques, so that we can have systems like JUPITER [3] capable of solving a domain-limited task whilst interacting with the user.

On the other hand, with a continuously growing explosion of information available on the World-Wide Web (WWW), an attractive database resource [4, 5], QA is a compelling framework for finding information that closely matches user's needs, aiming to retrieve answers instead of documents. In order to successfully match those user's needs, the QA has 'to understand' to a certain degree

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3214, pp. 131–138, 2004.

what the user is looking for. A typical QA performs several tasks that lead up to the 'user's question understanding' and therefore lead to selection of the best 'possible-answer' to the user request. Four of those tasks are worth mentioning since they possess a high level of importance: (a) question classification task, (b) query formation task, (c) document retrieval task (from the system knowledge database) and (d) answer selection task. Previous research [6, 7] has focused on the question classification task, prompting out its importance when selecting the answer since it tends to narrow the spectrum of possible-answer candidates. However, in spite of all the efforts, the task of extracting 'suitable answers' to the user request still remains barely solved.

In this paper we propose a Chat Environment for QA in open domain using WWW as the knowledge base. There are several goals we want to achieve with the Chat Environment. The main one is to acquire useful and precise information from the user for a better possible-answers selection. At the same time, we would like to achieve a user-computer interaction more like human-human interaction.

2 Basic Idea

We aim to achieve a QA system capable of holding a human-like interaction with the user. Fig.1 shows our system overview. As shown here the QA takes place within a Chat Environment (see the next section). The system processes

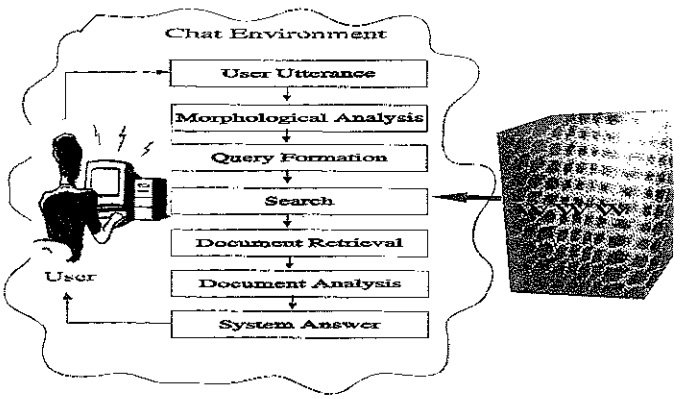

Fig. 1. Chat Environment for QA: System Overview

the user utterance using a morphological analyzer and sends a formed query to the Web search engine (using Google API [8]). The retrieved documents are processed in order to extract possible-answers to the user's request. A more detailed explanation of each process is described hereunder.

2.1 The Chat Environment (CE)

As a new point of view of the QA system in open domain, we started from this basis: in order to build the CE we created a 'chat-bot-like ELIZA-clone' oriented to obtain useful information from the user. A chat-bot is a computer program that simulates human chat. Chat-bots are wide spread over the Web and are used in applications such as electronic commerce (e-commerce) customer service, being their chats limited to specialized areas. With the use of an ELIZA-like chat-bot [9] improved with specific scripts oriented to obtain concrete information from the user about his/her request, we can better match the user's needs and at the same time achieve computer behavior (in terms of interaction) more like human behavior.

It is precisely the acquisition of information directly from the user that is one of the advantages that a QA in a CE could have over a typical QA, which just counts on the information extracted from the user's question. This *information acquisition* is particularly useful in cases where the information that can be obtained from a question is not enough for giving the user a suitable answer, as is the case of many "Who is (noun)?" or "Who are (noun)?" questions. The following chat example shows one of those cases:

System: Hi! How can I help you?
User: hi
System: How do you do? Please tell me your question.
User: yeah...
System: Pardon me? Did you say?
User: who is a diamond producer?
System: I'll search the Web. It may take a while and the possible-answers may be many. It'll help if you give me more information. What exactly are you looking for?
User: I'm looking for a diamond producer in America.
System: So you said "a diamond producer in America"?
.....

The chat is managed using scripts to simulate 'the system understanding' of the user utterances. In this case, the question "who is a diamond producer?" if treated by a typical QA, will retrieve from the Web around 340 thousand Web pages, with information about diamond producers all over the world, and since it is a "Who" question, most typical QA will search for a "person's name" as a possible-answer. With such an amount of information available, finding the most suitable answer to the user's request is very hard labor, in defiance of how good the question analysis done by a system could be. However, our QA, being in a CE, could acquire more specific information about what the user is trying to find. With the addition of this new information, "America", (see Query Formation section) the retrieval was diminished around 65% with information about 'Canada' as a diamond producer in *North* America, in the first Web page. Therefore, the CE helps to reduce considerably the number of Web pages to be processed in order to extract possible-answers. It is worth mentioning that another advantage of a QA in a CE can be found at the moment of analyzing

the question, because the question classification task could be avoided since the user herself/himself is providing the system with all the needed information for finding the possible sought-after answer.

2.2 Query Formation

When a *Web-question*[1] comes, the system performs a morphological analysis tagging the question. The system determines a Web-question by detecting question words (what, who, when, which, and so on; with exception of some questions about the system itself, like "What is your name?" or "How are you?"). We focus on dealing with factual questions. And more specifically, since 'who' questions tend to lack information (as shown in the chat example in the previous section), scripts were specifically designed to deal with them. The query is formed by extracting nouns, adjectives, adverbs and verbs (with some exceptions, like the verb To Be) from the given question. Thus, from the previous example:

Web-question detected
"who is a diamond producer?"
Morphological analysis

who	WP	who
is	VBZ	be
a	DT	a
diamond	NN	diamond
producer	NN	producer
?	SENT	?

Formed query (keywords)
diamond producer

Whilst the question is being analyzed, the system endeavors to obtain information from the user, in a chattering way. Hence, after the 'Web-question' is detected the following user utterance is morphologically analyzed for extraction of valuable information (nouns, mainly) to be added to the previously formed query. In our example, the new information obtained from the user is "America"(NN), then the new query becomes *diamond producer America*. This "augmented query" is then sent to the search engine and documents containing possible answers are retrieved.

2.3 Document Retrieval and Answer Selection

Since the documents retrieved from the Web are automatically ranked by the search engine according to their relevance regarding the query, and since the query formed within a CE contains user's precise information, the possible sought-after answers could be found within the first few documents. Therefore, our system just analyzes the first 20 HTML Web pages out of the thousands retrieved. The system parses the Web pages and segments each document into

[1] Web-question is a question whose answer is to be found searching the Web.

sentences. From those sentences, the ones selected as as possible-answers to be presented to the user, are extracted using the formula (1),

$$KeywordsinaSentence(KW_S) = \frac{n-1}{2} + 1 \quad (1)$$

where KW_S is a threshold and 'n' is the total number of keywords in the query. The sentences with KW_S or higher number of keywords are consider to be potential possible-answers since the minimum number of keywords in them is set to be more than a half of the total number of keywords in the query. Possible-answers extracted from the documents retrieved for our example (with the query: 'diamond producer America' -> n = 3; $KW_S = 2$) are:

– Canada: World's Third Largest **Diamond Producer**, Diamonds Net (Rapaport, January 4, 2004) According to a research paper released by Statistics...–
– Canada's diamond industry third-largest in world: Statistics Canada, OTT AWA- In just five years, Canada's burgeoning **diamond** business has put the country on track to become the third-largest **producer** in the world, Statistics Canada...–

3 Experiment and Results

One of the most notable differences between our QA in a CE and a typical QA (besides the smooth interaction with the user) is that the agent providing enough information for making it easier to find possible-answers is the user herself/himself; thus the question does not need to be rigidly classified. Hence, as we mentioned before, the question classification task could be avoided. In order to evaluate the effectiveness of the CE for QA, we compared the performance of our QA to the performance of a typical QA. Since in a typical QA the question needs to be classified, we created a probabilistic question classification system [10]. We will describe briefly our question classification process.

We defined question classification as the task that, given a question, selects from *n clusters* the one in which that question is more probable to appear. Those *n clusters* represent *n categories*. We assumed 24 clusters [ABB., ANIMAL, ART, BODY, COLOR, COUNTRY, CURRENCY, DATE, DEF., DESC., ENT., EXP., FOOD, GROUP, GEN.PLACE, MANNER, MED., PERC., PERSON, PROD., REASON, SUBS., SYN., TRANSP.] and for each one of them we built a first and second order Markov Model and combined those models using a linear combination. Since Markov Model suffers from sparseness, we extracted "valuable features" from each cluster. Those features are named entities, nouns and adjectives. They were ranked according to their frequency in each cluster as when a new question comes, its "valuable features" are extracted and the clusters where those features' frequency are high are the ones analyzed. In order to deal with the problem of unseen or unknown words that may appear in the test data, we used a combination of Back-off with Good-Turing smoothing technique [11]. As training data 3,865 questions from a corpus publicly available [12] were selected and distributed into the 24 clusters in order to build their Markov

Model. As test data 250 Text REtrieval Conference 10 (TREC 10) questions were distributed into 24 sets according to each cluster. This system achieved an accuracy of 81.3% classifying individual questions, and 21 out of 24 of the test data sets were correctly classified according to cluster or category, which means 91.6% accuracy for the classification of the sets. Once the question is classified, and the query is formed (using keywords from the question), documents are retrieved from the Web. Each category has answer patterns that are used for extracting possible-answers. For example, some answer patterns for the category abbreviation (ABB.) are:

(@1)?/NN/@2/NP*/@3/abbreviated/(possible-answer)/.
(@1)?/NP*/@2/NN/@3/acronimous/(possible-answer)/.
(@1)?/NN/stands for/(possible-answer)/.

Where @n represents possible text and NP, NN are proper nouns and nouns from the question. As for the question "what does NASA stand for?", classified correctly as ABB, sentences extracted as possible-answers were:

– NASA stands "for the benefit of all".–
– In the United States, NASA stands for the National Aeronautics and Space Administration.–

As we said before, with the experiment we tried to evaluate the effectiveness of a QA in a CE. We compared how well the possible-answers extracted by the CE-QA and the possible-answers extracted by the typical QA were related to the user's request. The sentences extracted as possible-answers were evaluated as Highly Related (HR), Related (R) or Barely Related (BR) to the user's sought-after answer according to their number of keywords *(kw)*. The possibility of No Extraction was contemplated as well. We selected 75 questions from corpora publicly available [12, 13]. Results are shown in Table 1.

From those results we can see that the CE-QA, as expected, could obtain better performance (around 85% of the sentences extracted as possible-answers were related to the user's sought-after answer) over a typical QA.

Table 1. Comparison between a Typical QA and CE-QA

System	HR ($kw > KW_S$)	R ($kw = KW_S$)	BR ($kw < KW_S$)	No Extraction
Typical QA	10%	60%	25%	5%
CE-QA	23%	67%	10%	-

4 Discussion

Preliminary experiments showed that a CE for a QA is effective for a more accurate possible-answer extraction. However, it is worth observing that, in spite of the huge amount of information available in the WWW, there were cases in which the systems' (both the typical QA and the CE-QA) performance was not good. For example, the question "who was the medieval classic hero that later

became the king of Denmark?" (from the corpus [13]) had No Extraction, using typical QA, and had the following BR sentences as possible-answers using the CE-QA:

- His (putative; Harald never recognized him) son Sweyn Forkbeard became King of Denmark, Norway and England.–
- Arthur, called the first 'worthy' of the Middle Ages, the British Charlemagne, famous in history, legend, and romance, became a renowned king in British History around whom an epic literature grew up over time, who, himself, evolved in medieval romance into the central figure of numerous tales about his knights, many of whom became celebrated figures themselves.–

We can see from this example that even though not always a suitable answer could be given to the user, a QA in a CE is always trying to find sentences that could match the user's needs. Thus, a CE for QA could be consider as a promising approach.

5 Conclusion

In this paper, we propose a simple CE for QA. Using a basic ELIZA-like CE as a promising approach to better match user's needs and at the same time to make a smoother user-computer interaction, we could see an improvement in the performance of our QA. Future works are oriented to widen the CE in order to deal with a bigger spectrum of questions that do not contain enough information as means to be suitably answered.

References

1. Weizenbaum J.: ELIZA-A Computer Program for the Study of Natural Language Communication Between Man and Machine. Communications of the ACM 9, No.1, pp. 36-45. (1966)
2. Colby K., Hilf F., Weber S.: Artificial Paranoia. Artificial Intelligence, Vol. 2, pp. 1-25. (1971)
3. Zue V. et al.: JUPITER: A Telephone-Based Conversational Interface for Weather Information. IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 1, pp. 85-96, January. (2000)
4. Kwok C., Etzioni O., Weld D.: Scaling Question Answering to the Web. In Proc. of the 10th International WWW Conference *(WWW10)*, pp. 150-161. (2001)
5. Chakrabarti S., van der Berg M., Dom B.: Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In Proc. of the 8th International WWW Conference *(WWW8)*, 1999. Also in Computer Networks, Vol. 31, No. 11-16, pp. 1623-1640. (1999)
6. Li X., Roth D.: Learning Question Classifier. In Proc. of the 19th International Conference on Computational Linguistics (COLING'02), pp. 556-562. (2002)
7. Zhang D., Lee W.: Question Classification Using Support Vector Machine. In Proc. of the 26th Annual International ACM SIGIR Conference, pp. 26-32. (2003)

8. Google API for Perl. Google Web APIs (beta) (2003).
http://www.google.com/apis/
9. Kimura Y., Araki K. et al.: Evaluation of Spoken Dialogue Processing Method Uisng Inductive Learning with Genetic Algorithm. In Proc. of the IASTED Int'l Conference Artificial Intelligence and Soft Computing, pp. 231-236. (2001)
10. Montero C., Araki K.: Probabilistic Question Classification. In Proc. of the 2004 IEICE General Conference, pp. 49, Tokyo Institute of Technology, Japan.
11. Li W.: Question Classification Using Language Modeling. Center of Intelligent Information Retrieval (CIIR). Technical Report. (2002)
http://ciir.cs.umass.edu/pubfiles/ir-259.pdf
12. Cognitive Computation Group at University of Illinois.
http://l2r.cs.uiuc.edu/~cogcomp/
13. Zheng Z.: AnswerBus Question Corpus Database. (2003)
http://134.96.68.36/corpus/answerbus.shtml