

# Accessing an Information System by Chatting

Bayan Abu Shawar and Eric Atwell

School of Computing, University of Leeds, LS2 9JT, Leeds, UK  
{bshawar, eric}@comp.leeds.ac.uk  
<http://www.comp.leeds.ac.uk/eric/demo.html>

**Abstract.** In this paper, we describe a new way to access information by “chatting” to an information source. This involves a chatbot, a program that emulates human conversation; the chatbot must be trainable with a text, to accept input and match it against the text to generate replies in the conversation. We have developed a Machine Learning approach to retrain the ALICE chatbot with a transcript of human dialogue, and used this to develop a range of chatbots conversing in different styles and languages. We adapted this chatbot-training program to the Qur’an, to allow users to learn from the Qur’an in a conversational information-access style. The process and results are illustrated in this paper.

## 1 Introduction

The widespread use of Internet and web pages necessitate an easy way to remotely access and search a large information system. Information extraction and information retrieval are about finding answers to specific questions. Information retrieval (IR) systems are concerned with retrieving a relevant subset of documents from a large set according to some query based on key word searching; information extraction is the process of extracting specific pieces of data from documents to fill a list of slots in a predefined templates.

However, sometimes we have less specific information access requirements. When we go to a conference and talk to research colleagues, we tend not to ask questions eliciting specific pointers to sources; rather, we chat more generally and we hope to get a bigger picture of general research ideas and directions, which may in turn inspire us to new ideas. An analogous model for information access software is a system which allows the information-seeker to “chat” with an online information source, to interact in natural language, and receive responses drawn from the online sources; the responses need not be direct “answers” to input “queries/questions” in the traditional sense, but should relate to my input sentences. The conversation is not a series of specific question-answering couplets, but a looser interaction, which should leave the user with an overall sense of the system’s perspective and ideas on the topics discussed. As a concrete example, if we want to learn about the ideas in the Qur’an, the holy book of Islam, then a traditional IR/IE system is not what we need. We may use IE to extract a series of specific formal facts, but to get an overview or broader feel for what the Qur’an teaches about broad topics, we need to talk around these topics with an expert on the Qur’an, or a conversational system which knows about the Qur’an.

In this paper we present a new tool to access an information system using chatting. Section 2 outlines the ALICE chatbot system, the AIML language used within it, and the machine learning techniques we used to learn Categories from a training corpus. In section 3 we describe the system architecture we implemented using the Qur'an corpus. We were able to generate a version of ALICE to speak like the Qur'an; this version and samples of chatting are discussed in section 4. Section 5 discusses the evaluation and the usefulness of such a tool. Section 6 presents some conclusions.

## 2 ALICE and Machine Learning Chatbots

In building natural language processing systems, most NLP researchers focus on modeling linguistic theory. [1] proposes an alternative behavioural approach: "Rather than attacking natural language interactions as a linguistic problem, we attacked it as a behavioural problem... The essential question is no longer "How does language work?" but rather, "What do people say?" ALICE [2], [3] is a chatbot system that implements human dialogues without deep linguistic analysis. The Alice architecture is composed of two parts: the chatbot engine and the language knowledge model. The chatbot engine is a program written in different languages, the one we used is written in Java, and is used to accept user input, search through the knowledge model and return the most appropriate answer.

The language knowledge model is based on manual authoring, which means the information model is hand crafted and it needs months to create a typical knowledge base. This knowledge is written using AIML (Artificial Intelligent Mark up Language), a version of XML, to represent the patterns and templates underlying these dialogues. The basic units of AIML objects are categories. Each category is a rule for matching an input and converting to an output, and consists of a pattern, which represents the user input, and a template, which implies the ALICE robot answer. The AIML pattern is simple, consisting only of words, spaces, and the wildcard symbols `_` and `*`. The words may consist of letters and numerals, but no other characters. Words are separated by a single space, and the wildcard characters function like words. The pattern language is case invariant.

Since the primary goal of chatbots is to mimic real human conversations, we developed a Java program that learns from dialogue corpora to generate AIML files in order to modify ALICE to behave like the corpus. Chatbots have so far been restricted to the linguistic knowledge that is "hand-coded" in their files. The chatbot must be trainable with a text, to accept input and match it against the text to generate replies in the conversation. We have worked with the ALICE open-source chatbot initiative: in the ALICE architecture, the "chatbot engine" and the "language knowledge model" are clearly separated, so that alternative language knowledge models can be plugged and played. We have techniques for developing conversational systems, or chatbots, to chat around a specific topic: the techniques involve Machine Learning from a training corpus of dialogue transcripts, so the resulting chatbot chats in the style of the training corpus [4], [5], [6]. For example, we have a range of different chatbots trained to chat like London teenagers, Afrikaans-speaking South Africans, etc by using text transcriptions of conversations by members of these groups. User input is

effectively used to search the training corpus for a nearest match, and the corresponding reply is output.

If, instead of using a dialogue transcript, we use written text as a training corpus, then the resulting chatbot should chat in the style of the source text. There are some additional complications; for example, the text is not a series of “turns” so the machine-learning program must decide how to divide the text into utterance-like chunks. Once these problems are solved, we should have a generic information access chatbot, which allows general, fuzzy information access by chatting to a source.

### 3 System Architecture of the Qur’an Chatbot

The Qur’an is the holy book of Islam, written in the classical Arabic form. The Qur’an consists of 114 sooras, which could be considered as chapters, grouped into 30 parts. Each soora consists of more than one ayyaa (sections). These ayyaas are sorted, and must be shown in the same sequence. In this chatbot version, we used the English/Arabic parallel corpora. The input is in English, and the output is ayyaas extracted from Qur’an in both English and Arabic. The program is composed of three subprograms. The first subprogram is used to build the frequency list from the local database. The second one is used to originate patterns and templates. And the last is used to rearrange the patterns and templates and generates the AIML files. In the following subsections we describe the three subprograms in detail.

#### 3.1 Program (1): Creating the Frequency List

The original English text format looks like:

```
THE UNITY, SINCERITY, ONENESS OF GOD, CHAPTER NO. 112
With the Name of Allah, the Merciful Benefactor, The Merciful Redeemer
112.001 Say: He is Allah, the One and Only;
112.002 Allah, the Eternal, Absolute;
```

The first line represents the soora title and number. The second line is the opening statement, which appears in every soora except soora number (9). The following lines hold the ayyaas, where each ayyaa has a unique number represented the soora number and the ayyaa number. After removing the numbers, a tokenization process is activated to calculate word frequencies.

#### 3.2 Program (2): Originating Patterns and Templates

Here we used the same program of generating other dialogue chatbots with some modifications. The Arabic corpus was added during this phase. Because of the huge size of the Qur’an text, we split the English/Arabic text into sub texts. Three processes are used within this program as follows:

1. Reading and concatenation process: the program starts by applying three reading processes; reading the English text with its corresponding Arabic one. Reading the English frequency list and inserting the words and its frequencies in least and count lists. Any ayyaa splits over two or more lines are merged together, so each element of the list represents a whole ayyaa.
2. Reiteration and finding the most significant word: since the Qur'an is not a series of "turns" so the machine-learning program must decide how to divide the text into utterance-like chunks. We proposed if an input was an ayyaa then the answer would be the next ayyaa in the same soora. During this process each element of the list except the opening line is repeated to be a pattern in one turn and as a template in the next. The list is now organized such that even indices holding the patterns while the odd ones holding the templates. A tokenization process is activated on each pattern. Then the frequency of each token is extracted from the count list produced in program (1). Finally the least frequent token is extracted.
3. Originating patterns/templates: after finding the least frequent word, the atomic and default categories are originated as follows; the atomic category will have the English pattern, which is the ayyaa, and its English template, which is the next ayyaa. The default category will have the least frequent word(s) as a pattern connected with "\*" in different position to match any word, and the template is the ayyaa(s) holding this word(s). During this phase, the English and Arabic soora numbers are replaced by the corresponding soora names. The Arabic template is appended to the English one. At the end, the generated patterns and templates are copied to a file.

### 3.3 Program (3): The Restructuring Process

Two lists are generated in this process: one for the atomic and another for the default categories. Given a large training corpus like Qur'an, where some ayyaas may be said more than once in the same soora or in different sooras, the file must be restructured to collate these repetitions. At the end the final categories are copied in AIML files.

## 4 Results

Before retraining ALICE with the generated AIML files, these files must be refashioned to enable ALICE interpreter to recognise the Arabic characters. This implies encoding the files using UTF-8 code. After all, two versions of ALICE were published using the Pandorobot service [8], the first named Qur'an0-30, which handles sooras from 1 to 30, and the second is Qur'an14-114, which handles sooras from 14-114. The program was able to generate 122,469 categories scattered in 45 files. Chatting (1) illustrates the sort of chat a user can have with our Qur'an chatbot.

## Chatting 1: chatting with Qur'an14-114

---

*Human:* Is it right that our protector is Allah?

*Qur'an14-114:* {Muhammad (Muhammad)} 11 That is because Allah is the Protector of those who believe, but those who reject Allah have no protector.

11(دمحم قروس) ذَلِكَ بَأْنُ اللّٰه مَوْلَى التّٰمِنِ اْمْتُوا وَاَنْ الكٰفِرِيْنَ لآ مَوْلَى لَهُمْ

{The Morning Star (At-Tariq)} 4 There is no soul but has a protector over it.

4(قراطل ا قروس) اِنْ كُلُّ نَفْسٍ لَّمَّا عَلَيْهَا حَافِظٌ

*Human:* Allah, the Eternal, Absolute

*Qur'an14-114:* {Sincerity (Al-Iklas)} 3 He begetteth not, nor is He begotten;

3(صالخال قروس) لَمْ يَلِدْ وَاَمْ يُولَدْ

*Human:* He begetteth not, nor is He begotten

*Qur'an14-114:* {Sincerity (Al-Iklas)} 4 And there is none like unto Him.

4(صالخال قروس) وَاَمْ يَكُنْ لَهُ كُفُوًا اَحَدٌ

The user types an utterance, such as a question or a statement; the system responds with one or more quotations (sooras and ayyaas) from the Qur'an, which seems appropriate. As this is a chat rather than accessing an information system, the ayyaas found are not simply the result of keyword-lookup; the response-generation mechanism is in fact hidden from the user, who will sometimes get the response "I have no answer for that".

## 5 Evaluation

Evaluation of this kind of general information access is not easy. As the information accessed is not in terms of specific questions, we cannot count numbers of "hits" in order to compute precision and recall scores. Up to now we evaluated chatting with Qur'an based on user satisfaction. We asked Muslims users to try it. And we asked them to answer the following questions: do you feel you have learnt anything about the teachings of the Qur'an? Do you think this might be a useful information access mechanism? If so, who for, what kinds of users? Some users found the tool unsatisfactory since it does not provide answers to the questions. However using chatting to access an information system, can give the user an overview of the Qur'an contents. It is not necessary that the user will have a correct answer for their request, but at least there is a motivation to engage in a long conversation based in using some of the outputs to know more about the Qur'an. Others found it interesting and useful in the case of a verse being read and the user wants to find out from which soora it came from. This would also benefit those who want to know more about the religion to learn what the Qur'an says in regards to certain circumstances, etc. They consider chatting with Qur'an as a searching engine with some scientific differences.

This tool could be useful to help students reciting Qur'an. All Muslims are taught to recite some of the Qur'an during school. However students usually get bored of the traditional teaching, such as repeating after the teacher or reading from the holy book. Since most students like playing with computers, and chatting with friends, this tool may encourage them to recite certain soora by entering an ayyaa each time. Since it is

a text communication, students must enter the ayyaa to get the next one, and this will improve their writing skills.

In previous research we evaluated our Java program depending on three metrics: dialogue efficiency, dialogue quality, and user satisfaction. From the dialogue efficiency and quality we aim to measure the success of our machine learning techniques. The dialogue efficiency measures the ability of the most significant word to find a match and give an answer. In order to measure the quality of each response, we classified the responses to three types: reasonable, weird but reasonable, or nonsensical. The third aspect is the user satisfaction. We applied this methodology on the Afrikaans dialogues [7], and we plan to apply it to the Qur'an chatbot as well.

## 6 Conclusions

This paper presents a novel way of accessing information from an online source, by having an informal chat. ALICE is a conversational agent that communicates with users using natural languages. However ALICE and most chatbot systems are restricted to the knowledge that is hand-coded in their files. We have developed a java program to read a text from a corpus and convert it to the AIML format used by ALICE. We selected the Qur'an to illustrate how to convert a written text to the AIML format to retrain ALICE, and how to adapt ALICE to learn from a text which is not a dialogue transcript. The Qur'an is the most widely known Arabic source text, used by all Muslims all over the world. It may be used as a search tool for ayyaas that hold same words but have different connotations, so learners of the Qur'an can extract different meaning from the context, it may be good to know the soora name of a certain verse. Students could use it as a new method to recite the Qur'an.

## References

1. Whalen, T.: Computational Behaviourism Applied to Natural Language, [online], <http://debra.dgrc.crc.ca/chat/chat.theory.html> (1996)
2. ALICE. A.L.I.C.E AI Foundation , <http://www.Alicebot.org/> (2002)
3. Abu Shawar, B. and Atwell, E.: A Comparison Between Alice and Elizabeth Chatbot Systems. School of Computing research report 2002.19, University of Leeds (2002)
4. Abu Shawar, B. and Atwell E.: Machine Learning from Dialogue Corpora to Generate Chatbots. Expert Update Journal, Vol. 6. No 3 (2003) 25-29.
5. Abu Shawar, B. and Atwell, E.: Using Dialogue Corpora to Train a Chatbot in: Archer, D, Rayson, P, Wilson, A & McEnery, T (eds.) Proceedings of CL2003: International Conference on Corpus Linguistics, Lancaster University (2003) 681-690.
6. Abu Shawar, B. and Atwell, E.: Using the Corpus of Spoken Afrikaans to Generate an Afrikaans Chatbot. to appear in: SALALS Journal of Southern African Linguistics and Applied Language Studies, (2003).
7. Abu Shawar, B. and Atwell, E.: Evaluation of Chatbot Information System. To appear in MCSEA'104 proceedings (2004).
8. Pandorobot: Pandorobot chatbot hosting service, <http://www.pandorabots.com/pandora> (2003)