ISBN 978-93-84422-37-0
2015 International Conference on Advances in Software, Control and Mechanical Engineering (ICSCME'2015)
Antalya (Turkey) Sept. 7-8, 2015 pp. 1-6

A chatbot as a Question Answering Tool

Bayan AbuShawar[1] and Eric Atwell[2]

[1]IT department; Arab Open University, Amman-Jordan;

[2]School of Computing; University of Leeds; Leeds UK

Abstract**:** *A chatbot is a program which can chat in natural language, on a topic built into the chatbot's internal knowledge model. Many chatbots exist, with different knowledge-bases programmed by the chatbot builders. We have built a system to convert a website text (corpus) to the chatbot format language. In this paper the chatbot is used as an interface for a Frequently-Asked Questions website. As an example, the FAQ website of the School of Computing in University of Leeds has been used to build FAQchat. Other FAQs from Leeds University have been investigated to extend the knowledge base of the chatbot and to investigate the use of the chatbot as a user-friendly interface to www information portals.*

Keywords:*Chatbot, FAQ, human computer interface, machine learning, corpus.*

1. Introduction

Question answering (QA) systems are developed to accept user's questions in natural language, and retrieve answers from either the Internet, or question-answer databases. The goal of the question answering system is "to retrieve 'answers' to questions rather than full documents or even best-matching passages as most information retrieval systems currently do" [14]. Several different QA systems have been built to retrieve answers for questions. For example, MULDER [16] only accept questions entered in the English language and then submits them to Google; whereas AnswerBus[22] allows users to enter questions in German, French, Italian, and Portuguese , then translates them to English and submits them to more than one search Engine, HealthQA [21] which is a Chinese QA system for smart Health, AQuASys [10]. Instead of retrieving answers from the web, FAQ-Finder [18] uses files of frequently asked questions extracted from USENET News as its knowledge base, Athira et al., [9] describes an architecture of a semantic web question answering based on ontological information. However, the retrieving process for this is not that simple, as these systems use sophisticated language processing to analyse the user input and retrieve answers by applying grammar and semantic parsers.

To avoid complexity in dealing with user's question, we use the chatbot as a tool to access information without the need for semantic or morphological analysis. A chatbot is a conversational agent that interacts with users using natural language. Originally the chatbot developers aim was to fool users that they were talking with real human. The first chatbot ELIZA emulated a psychotherapist [20], and then Colby [12] developed PARRY to simulate a paranoid patient. "Colby regarded PARRY as a tool to study the nature of paranoia, and considered ELIZA as a potential clinical agent who could, within a time-sharing framework, autonomously handle several hundred patients an hour." [15].

Nowadays several chatbots are available online, and are used for different purposes such as: MIA [1] which is a German advisor on opening a bank account; Sanelma [17] a fictional female to talk with in a museum that provides information related to specific piece of art; Cybelle [13], and AskJevees [8], a web- based search engine. Each of these commercial chatbots had to be "hand-trained" with question-patterns and answers for the specific domain; and the systems are not available for other researchers.

Instead of being restricted to a certain domain or written language, a Java program was developed to convert a machine readable text to the AIML format used by ALICE. We have worked with the ALICE open-source chatbot initiative. ALICE [19] is the Artificial Linguistic Internet Computer Entity, originated by Wallace in 1995. 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. Another major difference

between the ALICE approach and other chatbot-agents such as AskJeeves can be seen in the deliberate simplicity of the pattern-matching algorithms: whereas AskJeeves uses sophisticated Natural Language Processing techniques including morphosyntactic analysis, parsing, and semantic structural analysis, ALICE relies on a very large number of basic "categories" or rules matching input patterns to output templates. ALICE goes for size over sophistication: it makes up for lack of morphological, syntactic and semantic NLP modules by having a very large number of simple rules. The default ALICE system comes with about fifty thousand categories, and we have developed larger versions, up to over a million categories or rules.

We have techniques for developing new ALICE language models which can chat around a specific topic: these techniques involve machine learning from a training corpus of dialogue transcripts, so the resulting chatbot chats in the style of the training corpus [2], [3], [4], [5], [7]. For example, we have a range of different chatbots trained to chat like London teenagers, Afrikaans-speaking South Africans, loudmouth Irishmen, etc by using text transcriptions of conversations by members of these groups. The training corpus is in effect transformed into a large number of categories or pattern- template pairs. User input is used to search the categories extracted from the training corpus for the nearest match, and the corresponding reply is output.

We adapted our chatbot-training program to the FAQ in the School of Computing (SoC) at the University of Leeds, producing the FAQchat system. The replies from FAQchat look like results-pages generated by search engines such as Google, where the outcomes are links to exact or nearest match web pages. A search engine is "a program that searches documents for specific keywords and returns a list of the documents where the keywords were found." [11]. However FAQchat could also give a direct answer, if only one document matched the query; and the algorithm underlying each tool is different.

In this paper, section 2 describes the School of Computing FAQ Website. The Java program that processed the FAQ website is presented in section 3. Section 4 discusses the evaluation of FAQchat. Section 5 shows the possibility of extending the knowledge-base of FAQchat. Section 6 presents the conclusion that the chatbot could be used as an interface to www FAQ pages, as an alternative to other front-ends.

2. The School Of Computing Faq Website

The Frequently Asked Questions or FAQ website of the School of Computing is a structured database; thus most of the "data-cleaning" problems found with analysis of spoken dialogue corpora such as overlapping, and more than two speakers, are not found in the FAQs. Moreover almost all HTML tags are recognised by the ALICE interpreter because the AIML definition allows HTML tags to be embedded within templates. The questions and answers were extracted from the HTML files of the FAQ. The following cases, as presented in figure 2, need to be taken into consideration when adapting the java program to deal with the file format used in the School of Computing FAQ website:

- Each file has a title, and an interface to scroll up and down in the page during navigation, as shown in sample 1.a. The interface is not necessary and can be treated as redundant annotations .
- Some questions are marked by: <DIV CLASS="sect1"> tag and the answers by: <p> as shown in sample 1.b. The problem is when to consider the <p> tag as a part of the question, or when it denotes the beginning of the answer.
- Another problem illustrated in sample 1.b. is the reference problem, such as in the question "What is it?": "it" refers to what?
- Some questions are marked by: <DIV CLASS="question"> tag and the answer by: <DIV CLASS="answer"> as shown in sample 1.c.
- Using special character entities denoted by the "&" sign; these are not allowed in AIML.

Sample 1.a The extra notations used within FAQ files

```
<TR><TD WIDTH="10%" ALIGN="left" VALIGN="bottom">
<A HREF="appa02.html" ACCESSKEY="P">Prev</A></TD>
<TD WIDTH="80%" ALIGN="center" VALIGN="bottom">Maintenance of the FAQ</TD>
<TD WIDTH="10%" ALIGN="right" VALIGN="bottom">
```

<A HREF="appa04.html" ACCESSKEY="N">Next</A></TD></TR>

Sample 1.b Questions-answers denoted by <"sect1>", and "<p>" tags

<DIV CLASS="sect1">

<H1 CLASS="sect1">

<A NAME="x0101"> </A>What is it?</H1>

<P>The <SPAN CLASS="acronym">FAQ

</SPAN> contains information, mainly in question and answer form, for students and staff on many aspects of the School of Computing, concentrating on the School's computing facilities.

</P> <P>It contains advice on: </P><P></P>

<UL><LI>

<P><A HREF="x02.html">basic computer usage </A> (logging in, changing passwords) </P></LI> <LI>

<P>location and use of <A HREF="x09.html">computer laboratories</A> </P></LI>

Sample 1.c Questions-answers denoted by "question" and "answer" tags

<DIV CLASS="question">

<P> <A NAME="AEN598"></A>

<B>1. </B>

<SPAN CLASS="bold"><B CLASS="emphasis">What rules are there for the use of computing facilities?

</B> </SPAN> </P></DIV>

<DIV CLASS="answer">

<P> <B> </B>The School has a <A HREF="aup.html">Policy on the acceptable use of computing facilities</A>. All persons wishing to use the School's computing facilities will be required to read, agree to, and sign the acceptable use policy.</P> </DIV>

Fig 2.: FAQ sample and cases as found in SoC HTML files

3. Processing The Faq Website

We developed a java program to convert a readable text (corpus) to the AIML format. The program was tested with different corpora. In this paper we describe the version that handles the FAQs. The program is composed of four sub-programs as follows:

Sub-program 1: Reading the links and constructing a file of all links.

Sub-program 2: Generating the atomic file by reading questions and answers.

Sub-program 3: Constructing the frequency list, and a file of all questions.

Sub-program 4: Generating default files.

Sub-program 1: Creating links file: The FAQ is read to extract all links and put them in a file after eliminating any repeated links.

Generating Atomic file: The second program is for generating the atomic file; during this program the following modules are applied:

a. Extracting questions

1. Reading the questions which are denoted by specific tags illustrated in figure 3.13 such as <DIV CLASS=\"question\"> and <H1 CLASS="sect1">.
2. Concatenating the question lines until is encountered.
3. Normalising the question by removing punctuations, and un-necessary tags.
4. Adding the question as a pattern.

b. Extracting the answer

1. Reading the answer which is denoted by: <DIV CLASS=\"answer\">.

2. Checking that the number of the <Div class..> tags are equal to the number of the that denote the end of tags. If the number is not equal, the extra tags will be removed. This module was necessary to avoid the incompatible number of begin and end tags.
3. Replacing special character entities starting with "&" with normal alphabetic character.
4. Extracting the link for each question from the links file and adding it at the end of the template preceded by "For more information look at:"

Generating the frequency list

The frequency list is from the questions only, since the most significant words will be used within the questions. All questions denoted by <pattern> are read form the atomic file. The frequency is calculated using the same module as in previous prototypes.

Generating the default file

1. Reading the questions and extracting the two most significant words (content words only) which are the least frequent words.
2. Extracting the links that involve the most significant words.
3. Different categories are added to extend the chance of finding answers, where the answer is either a set of links or a direct answer as shown below:

- Build four categories using the most significant word (least 1) in four positions as patterns and the set of links it has as templates.
- Repeat the same using the second-most significant word (least 2)
- Build four categories using the first word and the most significant words (least 1) where the most significant word is handled in four positions.
- Build two categories using most significant 1 and most significant 2, keeping the order of position as in the original question. The answer is the set of links having both words, or if it is only one link, then the answer will be mapped to the pattern.
- Build a category using the first word, most significant word 1, and most significant word2 where the template is a direct answer.

At the end a version of ALICE called FAQchat [6] was generated to give answers to question relating to the School of Computing at University of Leeds.

4. FAQ Chat Evaluation

A comparison was carried out between FAQchat and Google search engine. Questions related to the School of Computing were provided to both tools FAQchat and Google. In our user trials, feedback favourable to FAQchat was gained from almost all users, even those who preferred Google. They found it a novel and interesting way to access the FAQ using natural language questions. Overall, about two thirds of users managed to find answers by FAQchat, and about two thirds of the users preferred to use it.

The number of evaluators who managed to find answers by FAQchat and Google was counted, for each question. Results in table 1 shows that 68% overall of our sample of users managed to find answers using the FAQchat while 46% found it by Google. 51% of the staff, 41% of the students, and 47% overall preferred using FAQchat against 11% who preferred the Google.

TABLE I: Proportion of users finding answers

Users /Tool	Mean of users finding answers		Proportion of finding answers	
	FAQchat	Google	FAQchat	Google
Staff	5.53	3.87	61%	43%
Student	8.8	5.87	73%	49%
Overall	14.3	9.73	68%	46%

Both staff and students preferred using the FAQ chat for two main reasons:

a. The ability to give direct answers sometimes where Google was only able to give links.
b. The number of links returned by the FAQ chat was less than those returned by Google for some questions, which saved time browsing/searching.

5. Extending The Knowledge-Base Of FAQ Chat

In order to enlarge the FAQchat knowledge base and to build a guide tool for students inside the campus of Leeds University to enrich their knowledge and receive immediate responses to questions that arise during their studies, various FAQs from different departments were investigated to train FAQchat. Our big target is to use the FAQchat as a tool to support e-learning. Distance e-learning is currently a matter of discussion and many systems have been implemented that utilise the power of technology and the World Wide Web to produce virtual classrooms. FAQ systems offer an asynchronous model that students can use to search previous lessons that have been posted by a teacher to find relevant answers. That means that there is no need for special outfitted classrooms and expensive equipment and students can work at their own pace.

The FAQ documents should have the same structure in order for the users to easily navigate them. At the moment there is no strict rule about how these must be constructed. However, the FAQ writers usually prefer to copy the structure of other publicly known sites in order to exploit past experience and create efficient and well-formed FAQs. Investigating the FAQ of other departments in Leeds University, we found that 63% create FAQs with multiple pages, 82% use links at the beginning of the FAQ to facilitate the navigation, and 68% use the "back to top" method to achieve quick navigation and transferability.

6. Conclusions

In this paper, we described a way to access information using a chatbot, without the need for sophisticated natural language processing or logical inference. FAQs are Frequently-Asked Questions documents, designed to capture the logical ontology of a given domain. Any Natural Language interface to an FAQ is constrained to reply with the given Answers, so there is no need for deep analysis or logical inference to map user input questions onto this logical ontology. To test this hypothesis, the FAQ in the School of Computing at the University of Leeds was used to retrain the ALICE chatbot system, producing FAQchat. The replies from FAQchat looked like results generated by search engines such as Google.

In our user trials, feedback favourable to FAQchat was gained from almost all users, even those who preferred Google. They found it a novel and interesting way to access the FAQ using natural language questions. Overall, about two thirds of users managed to find answers by FAQchat, and about two thirds of the users preferred to use it. The aim was not to try to come up with relative scores for the two systems, but to show that an ALICE- style corpus-trained chatbot is a viable alternative to Google and it could be used as a tool to access FAQ databases.

In order to extend the knowledge base of the FAQchat, different FAQs from different departments in Leeds University were investigated. These FAQs had different annotated format, which meant different normalisation processes were needed for each FAQ. This in turn necessitated the use of standard structure to generate FAQ web pages.

We managed to demonstrate that a simple ALICE-style chatbot engine could produce results at least as well-appreciated as those from the most popular commercial web search-engine. We did not need sophisticated natural language analysis or logical inference; a simple (but large) set of pattern-template matching rules was sufficient. We will make our Java tools available to others for research use.

Maybe it’s time EVERY information portal website got a chatbot!

7. References

[1] Aitools.org. (2004). [Online]: http://www.aitools.org/livebots/

[2] Abu Shawar, B., Atwell, E. (2003a). Using dialogue corpora to train a chatbot in: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) Proceedings of CL200, pp.681-690

[3] Abu Shawar., B., Atwell, E. (2003b). Machine learning from dialogue corpora to generate chatbots. Expert Update, vol. 6, pp. 25-30.

[4] Abu Shawar, B., Atwell, E. (2003c). Using the corpus of Spoken Afrikaans to generate an Afrikaans chatbot. Southern African Linguistics and Applied Language Studies. Vol. 21, pp. 283-294.
http://dx.doi.org/10.2989/16073610309486349

[5] Abu Shawar, B., Atwell, E. (2004). An Arabic chatbot giving answers from the Qur'an. In: Bel, B & Marlien, I (editors) Proceedings of TALN04. Vol 2, pp. 197-202 ATALA.

[6] Abu Shawar Bayan, Atwell Eric and Roberts Andy. (2005b). FAQchat as an information retrieval system. In: Zygmunt V. (ed.), Human Language Technologies as a Challenge for Computer Science and Linguistics: Proceedings of the 2ndLanguage and Technology Conference, Wydawnictwo Poznanskie, Poznan, pp. 274-278.

[7] Abu Shawar, B. Atwell, E. (2005a). Using corpora in machin- learning chatbot systems. International Journal of Corpus Linguistics 10:4, pp. 489-516
http://dx.doi.org/10.1075/ijcl.10.4.06sha

[8] AskJevees. (2004). [Online]: http://ask.co.uk/home

[9] Athira P., Sreeja M. and P. Reghuraj. 2013. Architecture of an Ontology-Based Domain-Specific Natural Language Question Answering System. International Journal of Wen & Semantic Technology (IJWest) vol.4, No. 4. Pp.31-39
http://dx.doi.org/10.5121/ijwest.2013.4403

[10] Bekhti, S.; Rehman, A.; Al-Harbi,M.; Saba, T.: AQuASys an Arabic question-answering system based on extensive question analysis and answer relevance scoring. Inf. Comput. Int. J. Acad. Res.3(4), 45–54 (2011)

[11] Boyle, R. (2003). "Understanding search engines". COMP1600: SY11 Introduction to Computer Systems 1. Lecture Notes, School of Computing, University of Leeds. pp65-72.

[12] Colby, K. (1999). Human-computer conversation in a cognitive therapy program. In Wilks, Y. (eds.) Machine conversations. Kluwer, Boston/Drdrecht/London. Pp. 9-19.
http://dx.doi.org/10.1007/978-1-4757-5687-6_3

[13] Cybelle. (2004). [Online]: AgentLand.com

[14] Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A. (2002). Web question answering: is more always better?. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (SIGIR 2002). Tempere, Finland, pp. 291-298.
http://dx.doi.org/10.1145/564376.564428

[15] Güzeldere, G. and Franchi, S. (1995). Dilaogue with colourful personalities of early ai". In Constructions of the Mind, SEHR, volume 4, issue 2. [On line]: http://www.stanford.edu/group/SHR/4-2/text/toc.html

[16] Kwok, C., Etzioni, O., and S.Weld, D. (2001). Scaling question answering to the web. ACM Transcations on Information Systems. Vol. 19, No.3, pp. 242-262.
http://dx.doi.org/10.1145/502115.502117

[17] MUMMI (2004). A Concept for chatbot:"Sanelma" building engaging relationship between the work of art and the exhibition visitor. [Online]: http://www.mlab.uiah.fi/mummi/sanelma/

[18] Robin, B., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N. and Schoenberg, S. (1997). Question answering from frequently-asked question files: experiences with the FAQ finder system. AI magazine. Vol. 18, No. 2, pp. 57-66.

[19] Wallace, R. (2003) The elements of AIML style. ALICE AI Foundation.

[20] Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine, Communications of the ACM, Vol. 10, No. 8, pp36-45.
http://dx.doi.org/10.1145/365153.365168

[21] Yanshen Yin, Yong Zhang, Xiao Liu, Yan Zhang, Chunxiao Xing, Hsinchun Chen: HealthQA: A Chinese QA Summary System for Smart Health. Lecture Notes in Computer Science Volume 8549, 2014, pp 51-62

[22] Zheng, Z., AnswerBus Question Answering System. In Proceedings of the Conference on Human Language Technology, 2002b.
http://dx.doi.org/10.3115/1289189.1289238