JEP-TALN 2004 - session on Arabic Language Processing
An Arabic Chatbot Giving Answers from the Qur'an
Un Chatbot Arabe qui Donne des Réponses du Coran

Bayan Abu Shawar and Eric Atwell
School of Computing,University of Leeds, Leeds LS2 9JT, England
bshawar@comp.leeds.ac.uk and eric@comp.leeds.ac.uk

Résumé – Abstract

Nous présentons des techniques d'apprentissage automatique que nous utilisons pour produire un chatbot arabe, qui accepte de l'input en arabe et qui produit des réponses tirées du Coran. Un système qui apprend des modèles conversationnels d'un corpus de conversation transcrite a été utilisé pour produire une gamme de chatbots qui parlent des langues diverses, y compris l'anglais, le français et l'afrikaans. Nous considérons des aspects de la langue arabe qui posent des problèmes pour l'apprentissage du chatbot et nous discutons le processus révisé pour la manipulation du texte de formation, de l'input et de la sortie arabes. En principe, le Coran offre des conseils et des réponses à des questions religieuses et à d'autres thèmes. Nous avons donc utilisé le Coran comme corpus de formation pour notre chatbot. Comme le Coran n'est pas une transcription d'une conversation, nous avons adapté le processus d'étude pour que le chatbot puisse se charger de la structure du Coran en ce qui concerne les sooras et les ayyas. Le système qui s'ensuit accepte de l'input en arabe, et répond avec des ayyas pertinents du Coran.

We present machine-learning techniques used to generate an Arabic chatbot, which accepts user input in Arabic and generates replies extracted from the Qur'an. A system to learn conversational patterns from a Corpus of transcribed conversation has been used to generate a range of chatbots speaking different languages including English, French and Afrikaans. We review aspects of the Arabic language, which pose problems for chatbot-learning, and we discuss the revised process to handle Arabic training text and input/output. In principle, the Qur'an provides guidance and answers to religious and other questions; so we used the Qur'an as a training corpus for our chatbot. As the Qur'an is not a transcription of a conversation, we adapted the learning process to cope with the structure of the Qur'an in terms of sooras and ayyas. The resulting system accepts user input in Arabic, and answers with appropriate ayyas from the Qur'an.

Keywords – Mots Clés – Key Words

Conversation, Apprentissage Automatique, Corpus, Langage de Balisage d'Intelligence Artificielle, Arabe, Coran.

Chat, Machine Learning, Corpus, Artificial Intelligence Markup Language, Arabic, Qur'an.

1 Introduction

A chatbot is a conversational agent that interacts with users turn by turn using natural language. Chatbots could serve in different domains such as: customer service, educational guidance, web site help, and for fun. ALICE (Abu shawar and Atwell 2002), the Artificial Linguistic Internet Computer Entity, is a chatbot system that implements various human dialogues, using AIML (Artificial Intelligent Markup 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.

However these dialogue patterns and templates are manually hand crafted in their files, which restricts the chatbot to language and domain supplied by the programmer. To generalize ALICE we developed a java program to convert a readable text (corpus) to the chatbot language model format. The program was able to create different AIML files using different training corpora in different languages (Abu Shawar and Atwell 2003c). This was first tested using the English-language Dialogue Diversity Corpus (DDC) (Mann 2002) to investigate the problems of utilizing a range of alternative dialogue corpora (see Abu Shawar and Atwell 2003a). The machine-learning was extended to learn patterns based on "most significant words", and this version was tested using an Afrikaans corpus (Van Rooy 2002), to generate an Afrikaans-speaking chatbot (Abu Shawar and Atwell 2003b). In this paper we show how we adapted the program to generate Arabic AIML files extracted from the Qur'an. Section 2 presents the modifications to meet the requirements of Arabic language and the non-conversational nature of the Qur'an. The resulting system accepts user input in Arabic, and answers with appropriate ayyas from the Qur'an, these results and conclusions are presented in sections 3 and 4.

2 Learning from an Arabic training corpus

The Arabic language has 28 consonants and three vowels: a, i, u that can be short or long. According to the vowels Arabic texts could be either a vowelled text such as the language of Qur'an or an unvowelled one used in media. The Qur'an text is available via the Internet; and in principle the Qur'an provides guidance and answers to religious and other questions; we selected the Qur'an as a training corpus for our chatbot to aid religious understanding and access to the Qur'an. As the Qur'an is not a transcription of a conversation, we adapted the learning process to cope with the structure of the Qur'an in terms of sooras and ayyas. The resulting system accepts user input in Arabic, and answers with appropriate ayyas from the Qur'an.

2.1 Learning from the Qur'an text

The Qur'an is the holy book of Islam, written in the Classical Arabic form. The Arabic language (both style-wise and content-wise) in the Qur'an is considered as one of its miracles. The Qur'an consists of 114 sooras, which could be considered as sections, grouped into 30 parts (chapters). Each soora consists of more than one Ayya (sentence). These ayyas are sorted, and must be shown in the same sequence. The AIML-learning system has been revised to handle the non-conversational nature of the Qur'an.

2.2 System architecture

We split the program to three parts. Part one creates the frequency list of the Qur'an, part two generates the original pattern and templates files, and part three apply the restructuring phase and generates the AIML files.

2.2.1 Reading the text

Each soora in the Qur'an has a title and a number, and consists of more than one ayya (section). In the online demo, we display the Arabic and English text of any example to facilitate reading for non-Arabic speakers. For instance the Qur'an text we used is as follows:

Arabic Soora: سورة الإخلاص (112)

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

قُلْ هُوَ اللَّهُ أَحَدٌ {112/1} اللَّهُ الصَّمَدُ {112/2} لَمْ يَلِدْ وَلَمْ يُولَدْ {112/3} وَلَمْ يَكُن لَّهُ كُفُوًا أَحَدٌ {112/4}

صفحة (604) ... http://www.islam4u.com .... القرآن الكريم ... موقع الاشعاع الاسلامي

The English translation soora is:

THE UNITY, SINCERITY, ONENESS OF GOD (112)

With the Name of Allah, the Merciful Benefactor, The Merciful Redeemer

Say: He is Allah, the One and Only {112/1} Allah, the Eternal, Absolute {112/2} He begetteth not, nor is He begotten {112/3} And there is none like unto Him {112/4}.

The first line holding the soora title and number, the second line is the opening statement in any soora, the third and fourth line display the ayyas. The curly brackets represent the end of the ayya (end of section) and keep the soora number as numerator and the ayya number as denominator. Finally the last line address the website where we captured this text and the page number where this soora found in the Qur'an.

2.2.2 Pre-processing stage

The Qur'an is a non-conversational text, so there is no overlapping, breaking, or other linguistic features of spontaneous speech to be filtered as in other dialogue corpora we treated. But we add a new procedure to split the soora into its ayyas, so each vector element holds one ayya, and during this process the curly brackets are removed. The vector looks like:

Arabic text	*English translation*
قُـلْ هُـوَ اللّـــهُ أَحَـدٌ 112/1	Say: He is Allah, the One and Only 112/1
اللّـــهُ الصَّـــمَدُ 112/2	Allah, the Eternal, Absolute 112/2
لَـمْ يَـلِـدْ وَلَـمْ يُـولَـدْ 112/3	He begetteth not, nor is He begotten 112/3
وَلَـمْ يَكُن لَّـهُ كُـفُـوًا أَحَـدٌ 112/4	And there is none like unto Him 112/4

We assumed that if an input is an ayya, then the reply will be the next ayya. So each ayya must be a pattern to guide us to the suffixed ayya, and a template for the previous one.

Each vector element is tokenised and compared with the generated frequency list to obtain the frequency for each word, then the obtained list is sorted in ascending order, the first word is picked as the least frequent word and the second word as the second least word. However it is not that easy as rise form the surface, before tokenisation, the punctuation must be removed, in the Qur'an text we remove anything between the curly brackets from all patterns.

Then creating categories procedure is called to generate the original pattern and template as atomic categories, and using the * to create the default categories using the least frequent word approach. Many default categories have been added to handle the position of the least frequent word. In case of one least word, we have four categories to be added, which hold the word itself, or the word in the beginning, middle, and last of an ayya. The template for the default categories will be the ayya, which holds these words. And before the template is generated, the soora number is replaced by the soora name.

For example if the least frequent word for the first line in the previous example is "أَحَـدٌ" then four categories are added, the atomic one is:

<category><pattern>أَحَـدٌ</pattern>

<template>1 (سورة الإخلاص) قُـلْ هُـوَ اللّـــهُ أَحَـد</template> </category>

The other three are having the same template but with different patterns:

<pattern>* أَحَـدٌ</pattern>, <pattern>أَحَـدٌ *</pattern>, and <pattern>* أَحَـدٌ *</pattern>

One difference between Arabic and other languages is the absence of upper or lower case letters, so there is no need to convert patterns to all-capitals. At the end of this process the patterns and templates are written in a text file.

2.3 Part three: the restructuring process

This procedure deals with repeated patterns and templates: <srai> tags are used with categories that have different patterns with the same template, and the random list (<random> tag) is used to group different templates with the same pattern. Because of the size of the Qur'an and the number of the AIML categories generated (76,404 chatbot rules, larger than the 50K original English ALICE rule-set), the Qur'an had to be split into several files to be processed separately and resulting AIML outputs were merged.

3. Results and evaluations

To test Alice retrained with an Arabic AIML file using the Pandorabot host service (Pandorabot 2003), two conditions must be applied into any uploaded file. First the XML declaration at the head of the file must be: <?xml version="1.0" encoding="utf-8"?> instead of <?xml version="1.0" encoding="ISO-8859-1"?> because ISO-8859-1 doesn't include any Arabic characters. The second is the text file must be encoded using utf-8 code. The contents of the file has been encoded using the MS-Word by saving it as an "Encoded text", then select the utf-8 code. The program generated 76,404 categories from the Qur'an. One potentially interesting achievement is if an Ayya is repeated in more than one soora, you can find the following ones in addition to the soora name that have this ayya; this may be useful for statistical analysis by Qur'an scholars and others. Feedbacks from Arabic evaluators find it difficult to enter the vowelled Arabic words. For example: in order to enter a word like (ALzaytoon), after every letter we type we have to press SHIFT key and the key of the vowel, which causes a lot of mistakes. Sometimes the consonant is doubled so users can miss the sign of the doubled consonants (Shadda). Another point, which does not satisfy users; is selected one item from the random list, in case the word(s) is repeated in more than one ayya, one is selected randomly. Also they noticed that not all words have an answer, and this is related to least frequent word technique we used, so just two words from each ayyas generated the categories.

To improve the system, we modified the restructuring process in part three, so instead of using a random list, we use a list (<ul>) tag, and in this case a list is shown as a response. This modification shows an improvement in users satisfaction.

To solve the main problem represented by vowelled letters, we will try in the future to remove vowels from the generated patterns to release users from using it.

The Example below shows chatting dialogue with the Qur'an, where the human enters a vowelled Arabic word(s), and the chatbot (Qur'an28-30) replies by finding the ayya(s) holding human input.

Example:

Human: أَحَـدٌ
Qur'an28-30: 1(سورة الإخلاص) قُـلْ هُـوَ اللَّـهُ أَحَـدٌ
26(سورة الفجر) وَلَـا يُـوثِـقُ وَثَـاقَـهُ أَحَـدٌ

Human: قُـلْ هُـوَ اللَّـهُ أَحَـدٌ
Qur'an28-30: 2(سورة الإخلاص) اللَّـهُ الصَّـمَدُ

Human: عَـابِـدُونَ أَنـتُم
Qur'an28-30: 3(سورة الكافرون) لَـا أَنـتُمْ عَـابِـدُونَ مَا أَعْـبُدُ وَ
5(سورة الكافرون) وَلَـا أَنـتُمْ عَـابِـدُونَ مَا أَعْـبُدُ

---

6. Conclusions

Communications with computers using natural languages is one of the essential concerns of researchers. 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 and to the natural language, which is written or spoken. We have developed a java program to read a text from a corpus and convert it to the AIML format used by ALICE. The program has succeeded in creating new versions of ALICE speaking different languages: English, French, and Afrikaans. In this paper we illustrate how to convert an Arabic text to the AIML format to retrain ALICE, and how to adapt ALICE to learn from a text, which is not strictly a dialogue transcript.

The Arabic Language has its own characteristic features, which make it different than other languages. We selected the Qur'an text to generate Arabic AIML catgeories, where the text is written and using the Classical Arabic form, which is diacritical, marked, to avoid the problem of morpohosyntactic analysis in ALICE. A significant difference between training corpora we have used previously and the Qur'an is the non-conversational nature of its text. The Qur'an is the most widely known and used Arabic text in the world, so we hope to find many users and uses for a system to converse with it. It may be used as a search tool for ayyas that hold same words but have different connotations, so learners of the Qur'an can extract different meaning from the context. We were able to use the Qur'an as a demonstration training set, generating 76,404 categories, a larger language model than the default English AIML model supplied with ALICE (about 50,000 categories). We were also able to use the Qur'an to demonstrate the flexibility of our machine learning approach to deal with different corpora regardless if it is conversational or not.

Références

Abu Shawar, B. and Atwell, E. (2002). A comparison between ALICE and Elizabeth chatbot systems. School of Computing research report 2002.19, University of Leeds.

Abu Shawar, B. and Atwell, E. (2003a). Using dialogue corpora to retrain a chatbot system. Proceedings of the Corpus Linguistics 2003 conference (CL2003), Lancaster University pp. 681-690.

Abu Shawar, B. and Atwell, E. (2003b). Using the Corpus of Spoken Afrikaans to generate an Afrikaans chatbot in SALALS Journal: Southern African Linguistics and Applied Language Studies

Abu Shawar, B. and Atwell, E. (2003c). Machine Learning from dialogue corpora to generate chatbots. Proceedings of BSC-AISIG'03, Nottingham Trent University

Mann, W (2002) Dialog Diversity Corpus http://www-rcf.usc.edu/~billmann/diversity/DDivers-site.htm

Pandorabot (2003). Pandorabot chatbot hosting service http://www.pandorabots.com/pandora

Qur'an28_30 (2004). http://www.pandorabots.com/pandora/talk?botid=94fdd8596e348b29

Van Rooy, B. (2002). Transkripsiehandleiding van die Korpus Gesproke Afrikaans. [Transcription Manual of the Corpus Spoken Afrikaans.] Potchefstroom: Potchefstroom University.