

# A Chatbot as a Novel Corpus Visualization Tool

Bayan Abu Shawar & Eric Atwell

School of Computing, University of Leeds, Leeds LS2 9JT, England

[bshawar@comp.leeds.ac.uk](mailto:bshawar@comp.leeds.ac.uk) and [eric@comp.leeds.ac.uk](mailto:eric@comp.leeds.ac.uk)

## Abstract

The classical way of viewing data set is using the visualization process, which maps the data from numerical or textual form to a visual representation that our mind can easily interpret such as: using graphical diagrams, charts, and geometric representation. In this paper we introduce a new idea to visualize a dialogue corpus using a chatbot interface tool. We developed a java program to convert a readable text (corpus) to AIML language to retrain ALICE. We use specific domains of the BNC spoken files to retrain ALICE, and visualise the data contents of these domains via chatting. Even that visualizing the corpus using a chatbot based on the chatting time and the size of the corpus, certain domains could be scanned easily if they are part of the current corpus or not. Our main conclusion is that it is possible to use the chatbot tool as a visualization process of a dialogue corpus, and to model different chatbot personalities.

## 1. Introduction

Language is “the primary vehicle by which people communicate and record information”. (Grishman, 1986). Human machine conversation is a new technology integrating different areas where the core is the language model, and the computational methodologies, to facilitate informal chat communication between users and computers using natural language. A related term to machine conversation is the chatbot, a conversational agent that interacts with users turn by turn using natural language. Different chatbots or human-computer dialogue systems have been developed either using spoken or text communication. Chatbots have been used in different domains such as: customer service, educational guidance, web site help, and for fun. However these chatbots are restricted to the knowledge that is manually “hand-coded” in their files and to the natural language, which is written or spoken.

Alice (Abu shawar & Atwell, 2002; 2003a) is one of these chatbots; in fact, Alice is a general architecture for building chatbots, using the AIML formalism (Artificial Intelligence Markup Language) to represent its knowledge. Alice and AIML were designed to allow non-programmers to develop chatbots: only the rules (or “categories”) need to be specified by the developer. This still requires a lot of human effort; the Alice Botmaster developed the current public-domain set of AIML files for Alice, a linguistic knowledge base of more than 50,000 rules, over several years. In order to retrain Alice, we developed a java program to read a dialogue from a corpus and convert it to the AIML linguistic knowledge representation format. We used a range of corpora to create several different experimental version of Alice, speaking different varieties of English, as well as French, Afrikaans, and a bilingual chatbot. Our machine learning techniques succeeded in learning a linguistic knowledge base of more than one million categories or rules, extracted from the British National Corpus (BNC) spoken transcriptions.

Section 2 presents ALICE, the AIML format and the pattern matching techniques used. A description of the automation process and machine learning techniques are clarified in section 3. Results of using BNC and sample of chatting visualizing the corpus contents are shown in section 4. Section 5 includes our conclusion; which refers to the possibility of using the chatbot as a visualizing tool.

## 2. ALICE and the AIML Language

A.L.I.C.E is the Artificial Linguistic Internet Computer Entity. Alice knowledge about English conversation patterns is stored in AIML files. AIML, or Artificial Intelligence Mark-up Language, is a derivative of Extensible Mark-up Language (XML). It was developed by Wallace and the Alicebot free software community during 1995-2000 to enable people to input dialogue pattern knowledge into chatbots based on the A.L.I.C.E. open-source software technology.

AIML (Abu Shawar and Atwell 2003a), describes a class of data objects called AIML objects and partially describes the behaviour of computer programs that process them. AIML objects are made up of units called topics and categories. Categories are the basic unit of knowledge in AIML. Each category is a rule for matching an input and converting to an output, and consists of a pattern that represents the user input, a template that implies the ALICE robot answer, and an optional context called “that” and “topic”. The <that> tag holds the robot’s last utterance, and the “topic” collects a group of categories together.

The AIML pattern language is simple, consisting only of words, spaces, and the wildcard symbols `_` and `*`. The words may consist of letters and numerals, but no other characters. The pattern language is case invariant. Words are separated by a single space, and the wildcard characters function like words. Based on the patterns we have two types of categories: an atomic category and a default one.

*Atomic Categories:* are those with patterns that does not have wildcards “\*” or “\_”.

Example (1):

```
<category><pattern>HELLO</pattern>      <template>Hi there!</template></category>
```

If the user input is: Hello, then the robot will answer will be: Hi there!

*Default categories:* are those with patterns has a wildcards “\*” or “\_”. These patterns resulted from a reduction process while the robot searching for the appropriate matching. These wildcards match any input but they are differ in their alphabetical order.

Example (2):

```
<category>
<pattern>HELLO      *</pattern>      <template>Hi there!</template>
</category>
```

In this case the “\*” will match the suffixed string after Hello and the answer will be: “Hi there!”

The template is the AIML response or reply. In its simplest form, the template consists of only plain, unmarked text. More generally, the template may also contain AIML tags which can save data, activate other programs, give conditional responses, and recursively call the pattern matcher to insert the responses from other categories. For instance categories with templates having <sr> or <sr> tags are called recursive categories, which recursively call the pattern matcher to insert the responses from other categories. <sr> and <sr>, which refers to simply recursive artificial intelligence and symbolic reduction are playing critical rules in: reducing complex grammatical forms to simpler ones; splitting an input into two or more subparts, and combines the responses to each; and dealing with synonyms by mapping different ways of saying the same thing to the same reply.

Example (3):

```
<category>
<pattern>HALO</pattern><template><sr>HELLO</sr>
i</template>
</category>
```

This example illustrates dealing with synonymy, if the input matches “HALO”, it will recursively call another atomic category holding “HELLO” as a pattern. If we used example (2), the answer for “HALO” will be “Hi there!”.

This pattern matching architecture is simpler than other chatbot dialogue models. The matcher tries to give you the best answer by finding the longest pattern match. Scanning the input word by word using depth-first search to find the longest match, if fails, then a backtracking will be used, until a match found, or in bad case matches “I have no answer for that”.

### 3. System Description

What the Alice dialogue-model lacks in sophistication is made up for by quantity of rules. Alice knowledge is easier to extend, by adding more categories; there are more than 50,000 categories in the current public-domain Alice “brain”, added by the Botmaster over several years. However all these categories are manually “hand-coded”, which is time-consuming, and restricts adaptation to new discourse-domains and new languages.

To overcome this problem we automated the process of building AIML knowledge. We developed a java program that converts a readable text (corpus) to the AIML format. The program started by reading the corpus and insert in a vector. Then each vector element is passed through reprocessing phases where all linguistic annotations such as overlapping, fillers, and other linguistic annotations are filtered. After that a converter module considers the first turn as a pattern and the second as a template. The restructuring module takes its turn by searching the patterns template vector to map all patterns with the same response to one form, and to transfer all repeated pattern with different templates to one pattern with a random list of different responses. Finally all these categories are copied in AIML files.

The problem we faced is how to originate default patterns and templates to extend the probability of finding an answer. To solve it we designed a machine learning technique, which extracts the least frequent word from

each pattern and builds the default categories. The idea behind this approach, is imitating the people behaviourism during conversation, where people usually reply according to most significant point (word) during their talk (Abu Shawar & Atwell, 2003b). To obtain the least frequent word, the program creates a local frequency list.

The program was tested using different languages and corpora (Abu Shawar & Atwell, 2003c). 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. User input is effectively used to search the training corpus for a nearest match, and the corresponding reply is output.

#### 3.1 The BNC Chatbot

The British National Corpus (BNC) is a collection of text samples amounting to over 100 million words, extracted from 4124 modern British English texts of all kinds, both spoken and written. The corpus is annotated using SGML (XML-like) mark-up, including CLAWS Part-of-Speech category of every word. All annotations are marked between <angle brackets>. The corpus is partitioned into two types: the spoken and the written transcripts. (Herring 1996) argues that computer mediated communication (CMC) “is typed, and hence like writing, but exchanges are often rapid and informal, and hence more like spoken conversation”, and (Grondelaers; Speelman & Geeraerts, 2003) describe of the language of Internet Relay Chat (IRC) as an example of “spoken language in written form”; so we decided to retrain Alice using the BNC spoken transcripts.

The spoken dialogue transcripts amount to 10 million words, and can be divided into two parts: a demographic part, involving transcriptions of spontaneous natural conversations between families, friends, and so forth, and the context-governed part, containing transcription recorded in educational, informative, business, leisure, institutional, and public events. Each corpus file starts with a long Header section, containing details of source, speakers, etc. In the transcript Body, the dialogue consists of a series of utterances or speaker-turns, marked at start and end by <u> and </u> tags. Each utterance tag also includes a speaker number (anonymised, eg F72PS002). Within a text sample, all sentences are tagged <s> and numbered; and each word is preceded with a CLAWS Part-of-Speech tag, e.g. ITJ = interjection, PUN = punctuation-mark, NP0 = singular proper name. An example of a sequence of two utterances is:

```
<u who=F72PS002>
<s n="32"><w ITJ>Hello<c PUN>.
</u>
<u who=PS000>
<s n="33"><w ITJ>Hello <w NP0>Donald<c PUN>.
</u>
```

The corresponding basic literal AIML is:

```
<category>
<pattern>HELLO</pattern>
<template>Hello Donald</template>
</category>
```

## 4. Results

1,153,129 categories were generated and separated in 801 files and 47 folders, which took up 320 MB of disc space. This number is 20 times bigger than any existing chatbot: the large public domain. AIML category-file developed over several years contains about 50,000 categories. Since all these categories are taken from different domains, we found it reasonable to make each special domain as a separate version. The following table shows these versions.

ALICE version	Atomic categories	Default categories	Total
Sport	3283	7913	11196
World affairs	3120	8756	11876
Travel	640	1636	2276
Media	1061	3126	4187
Food	503	1125	1628

**Table 1 ALICE BNC versions**

Also we generated versions speak like teenagers such as: Robin, Peter, Michael, Samantha, and Jessica. All these and others are found at (<http://www.comp.leeds.ac.uk/eric/>).

Samples of chatting are shown below.

### Chatting (1)

**Human:** Hi, do you like sport?

**Bot:** Hi there. No I'm not much into sports. I like solo sports like rollerblading and skiing though.

**Human:** Do you anything about football?

**Bot:** Any of several games played with an inflated leather ball by two teams on a field with goals at each end, the object being to get the ball across the opponents' goal.

**Human:** What about women's football?

**Bot:** I think football seems to, to bring out the worst in a lot of people and that's the part of it that I do not like

**Human:** Do you like it?

**Bot:** and er, I think until women's football take it on board all the way erm, they're not really gonna stand up with the men they have to become, referees, they have to become coaches, and more and more of these people erm, taking responsibility I go along and watch women's football and get really infuriated to see a man out there refereeing the game, and often in a very condescending manner! Instead of carding someone he'll put his arm round her and say Do not !It's not very nice to do that !Do not do that again!

**Human:** It is an enjoyable game .

**Sport:** Aha Aha It is a case is not it, that if you want to be a er a professional woman footballer, certainly there's more opportunities on the continent than there are in Britain, is, is is is that the case? Yes?

**Human:** I watch it regularly.

**Sport:** Because if football's on it has to be watched! It's like a a religion! Fi

**Human:** Can you admit how many hours do you watch football?

**Bot:** If played properly, it's poetry in motion! And as the lady down here said, if the ladies do not like it, switch it off! If I had the chance I'd watch it twenty four hours a day!

**Human:** What about basketball?

**Sport:** mm

From the previous chatting we can know that the corpus involves conversation about football games. It seems that the speaker like football games, and he encourages women to play football too. From the last utterance we obtained that there is nothing about basketball.

### Chatting (2):

**Human:** One day when I woke up I had a terrible pain in my muscles.

**BNC:** what happens is both muscles contract together fully So you ve got one muscle trying to pull your arm that way another muscle trying to pull your arm that way and your muscles become rigid and locked There h have been cases where the person s back muscles have pulled up so tight that their they ve formed an arch Their head has been in t touch with the bed their heels but the whole body s been arched like a bow with the contraction of these muscles.

**Human:** does this happen in your jaw?

**Bot:** No it s all over your body But it s called lockjaw because er one of the stages is that your jaw actually locks solid

**Human:** did they find a cure

**Bot:** Yes

**Human:** good, in the past it leads to death

**Bot:** Mm Well that s what I was going to come to There are now only two places in the country that actually deal with tetanus poisoning One is Oxford and the other is in Leeds Er

**Human:** Sir, what do they do in Scotland then?

**Bot:** You re flown down in helicopter to Leeds

**Human:** Any side effect for the cure?

**Bot:** No but I ll come onto some effects that it did have in a moment Now If you ve got tetanus to stop you breaking your back with these muscle contractions so that they can nurse you adequately they inject you with this stuff and you are paralysed until the tetanus has been fought off They will give you the antitoxin they will give you the vaccination and they will keep you breathing Cos what normally happens is with tetanus that the reason it kills you is that you stop breathing because your muscles that work your lungs the diaphragm the intercostal muscles between the ribs those seize up and you just stop breathing So if they can keep you breathing and get your body to produce the necessary antibodies then they can cure you But it may take three months so you are three months possibly on a ventilator What will happen is at regular intervals they ll take you off the ventilator inject you with the er antidote to the depolarizer and see if your muscles lock up If they do another injection tube back down and back on the life-support machine Now bear in mind that you may be up anything up to three months in a condition where you re completely and totally unable to move a muscle for yourself That s

In chatting (2) we can conclude that it is a conversation between an expert in Biology, or medicine, who are talking about jaw, and tetanus.

In fact to have a clear insight about the corpus content, this based on the time spent in chatting, longer is better. You can notice from the previous, that some answers are a respond to the human, but even in this case, you can know what is there in the corpus from the robot answers.

## 5. 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 used versions of ALICE generated using the BNC spoken corpus. Our main conclusion is that even if the chatbot cannot provide correct answer for an input, it can give the user an insight about the corpus contents. Good visualization depend on the chatting time, the longer is better.

## 6. References

- Abu Shawar, B. & Atwell, E. (2002). A comparison between Alice and Elizabeth chatbot systems. School of Computing research report 2002.19, University of Leeds.
- Abu Shawar, B. & Atwell, E. (2003a). Using dialogue corpora to train a chatbot. In Proceedings of CL2003: International Conference on Corpus Linguistics (pp. 681—690). Lancaster University.
- AbuShawar, B. & Atwell, E. (2003b). Machine Learning from dialogue corpora to generate chatbots. *Expert Update journal*, 6(3), 25--29.
- AbuShawar, B. & Atwell, E. (2003c). 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.
- Grishman, R. (1986). *Computational Linguistics: an introduction*. Cambridge University Press. 1-9.
- Grondelaers, S., Speelman, D., Geeraerts, D. (2003). A corpus-based approach to informality: the case of Internet chat . In Proceedings of the Corpus Linguistics 2003 conference (CL2003) (pp. 264). Lancaster, University UK.
- Herring, S. (1996). Introduction. In Herring S (ed) *Computer-mediated communication: linguistics, social and cross-cultural perspectives*. Amsterdam John Benjamins. pp1--10.