Article:

Abu Shawar, BA and Atwell, ES (2005) A chatbot system as a tool to animate a corpus. ICAME Journal: International Computer Archive of Modern and Medieval English Journal, 29. 5 - 24. ISSN 0801-5775

A chatbot system as a tool to animate a corpus

Bayan Abu Shawar and Eric Atwell
University of Leeds

1 Introduction

"Before there were computers, we could distinguish persons from non-persons on the basis of an ability to participate in conversations. But now, we have hybrids operating between person and non persons with whom we can talk in ordinary language." (Colby 1999). Human machine conversation as a technology integrates different areas where the core is language, and the computational methodologies facilitate 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 using text communication starting from ELIZA (Weizenbaum 1966) that simulates a psychotherapist, then PARRY (Colby 1973) which simulates 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." (Guzeldere and Franchi 1995).

Nowadays a lot of chatbots are available online and are used for different purposes such as: MIA (MIA 2004) which is a German advisor on opening a bank account; and Sanelma (MUMMI 2004), a fictional female to talk with in a museum that provides information related to specific pieces of art.

Practical applications and evaluation are key issues in Language Engineering: Cunningham (1999) characterises Language Engineering in terms of "…its focus on large-scale practical tasks and on quantitative evaluation of progress, and its willingness to embrace a diverse range of techniques". The Loebner prize competition (Loebner 2003) has been used to evaluate machine conversation chatbots. The Loebner Prize is a Turing test, which evaluates the ability of the machine to fool people that they are talking to a human. In essence, judges are allowed a short chat (10 to 15 minutes) with each chatbot, and asked to rank them in terms of “naturalness”.

Most chatbots are restricted to knowledge that is manually “hand coded” in their files, and to a specific natural language which is written or spoken. To overcome this problem and to generate chat which is closer to human language, we developed a program to convert a machine readable text (corpus) to a specific chatbot format, which is used to retrain the chatbot.

Sinclair (1991) defined the corpus as “A collection of naturally occurring language text, chosen to characterize a state or variety of a language.” Linguists collect texts whether written or spoken to aid study of language features such as syntax, semantics and pragmatics. Leech (1992) states “The focus of study is on performance rather than competence and on observation of language in use leading to theory rather than vice versa.”. In other words, the collected text represents the corpus, and studying the language using a corpus based approach is known as corpus linguistics. With the rapid evolution of computer capacities and capabilities, many corpora have become available in machine readable form, and a lot of software tools have been developed to annotate the corpora, e.g. CLAWS used to PoS_tag the LOB and BNC corpora (Leech et al. 1994); and retrieve data from a corpus, e.g. XAIRA, a web based concordance application, developed for use with the British National Corpus (BNC) (Burnard and Dodd 2003); or lemmatised and unlemmatised frequency lists generated by Kilgarriff (1996).

In this paper, we present the chatbot system as a tool to explore or visualize different types of English language used in the BNC corpus in a qualitative manner in contrast to tools such as Wmatrix which visualises a corpus in terms of quantitative statistics. Section 2 reviews the ALICE chatbot system, the linguistic knowledge representation format and the pattern matching technique. In section 3, the Wmatrix tool is used to show the difference between human to human dialogues and ALICE chatbot dialogues. Section 4 presents the BNC corpus and the problems which arose during the automation process, and the software tool that maps the BNC spoken transcripts to ALICE chatbot internal knowledge representation, called AIML. Sample of dialogues generated from the BNC are shown in section 5. Section 6 presents our conclusions.

2 The ALICE chatbot system

“The need of conversational agents has become acute with the widespread use of personal machines with wish to communicate and the desire of their makers to provide natural language interfaces” (Wilks 1999).

ALICE (Alice 2000, Abu Shawar and Atwell 2002, 2003, Wallace 2003) is the Artificial Linguistic Internet Computer Entity, first implemented by Wallace in 1995. 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 consists of data objects called AIML objects, which are made up of units called topics and categories. The topic is an optional top-level element; it has a name attribute and a set of categories related to that topic. Categories are the basic units of knowledge in AIML. Each category is a rule for matching an input and converting to an output, and consists of a pattern, which matches against the user input, and a template, which is used in generating the Alice chatbot 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. The idea of the pattern matching technique is based on finding the best, longest, pattern match.

2.1 Types of ALICE/AIML categories

There are three types of categories: atomic categories, default categories, and recursive categories.

a. *Atomic categories* are those with patterns that do not have wildcard symbols, _ and *, e.g.:

   <category><pattern>HELLO ALICE</pattern>
   <template>hello</template></category>

   In the above category, if the user inputs ‘hello Alice’, then ALICE answers ‘hello’.

b. *Default categories* are those with patterns having wildcard symbols * or _. The wildcard symbols match any input but can differ in their alphabetical order. For example, given input ‘hello robot’, if ALICE does not find a category with exact matching atomic pattern, then it will try to find a category with a default pattern such as:

```
<category><pattern>HELLO *</pattern>
<template>hello friend</template> </category>
```

So ALICE answers ‘hello friend’.

c. *Recursive categories* are those with templates having <srai> and <sr> tags, which refer to simply recursive artificial intelligence and symbolic reduction. Recursive categories have many applications: symbolic reduction that reduces complex grammatical forms to simpler ones; divide and conquer that splits 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.

c.1 Symbolic reduction

```
<category> <pattern>DO YOU KNOW WHAT THE * IS</pattern>
<template> <srai>What is <star/></srai> </template> </category>
```

In this example <srai> is used to reduce the input to simpler form “what is *”.

c.2 Divide and conquer

```
<category> <pattern>YES*</pattern>
<template> <srai>YES</srai><sr/> <template> </category>
```

The input is partitioned into two parts, “yes” and the second part; * is matched with the <sr/> tag.

```
<sr/> = <srai><star/></srai>
```

c.3 Synonyms

```
<category> <pattern>HI</pattern>
<template> <srai>Hello</srai> </template> </category>
```

The input is mapped to another form, which has the same meaning.

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 "hand-coded", which is time-consuming, and restricts adaptation to new discourse-domains and new languages. In the following sections we will present the automation process we developed to train Alice using a corpus based approach.

3 Human to human versus human to chatbot dialogues

Before training ALICE-style chatbots with human dialogue corpus texts, we investigated the differences between human-chatbot dialogue and human-human dialogue (Abu Shawar and Atwell 2003a). To do this, we compared a dialogue transcript generated via chatting with ALICE, and real conversations extracted from different dialogue corpuses. The comparison illustrates the strengths and weaknesses of ALICE as a human simulation, according to linguistic features: lexical, Part-of-Speech, and semantic differences. The Wmatrix tool (Rayson 2002) was used for this comparison. Wmatrix computes Part-of-Speech class and semantic class for each word in the texts, and then highlights specific words, Part-of-Speech categories, and semantic word-classes, which appear more often in one text than the other. The comparison results are viewed as feature frequency lists ordered by log-likelihood ratio: highest LL values indicate the most important differences between corpora. We used Wmatrix to compare between human-to-human dialogues extracted from several sub-corpora included in the DDC, Dialogue Diversity Corpus (Mann 2002), and human-to-computer dialogues extracted from chats with ALICE on the AI movie website (Spielberg 2000). Four different corpuses in different fields and sizes were investigated; the DDC sub-corpora and ALICE transcript are not equal in size, so we look at the relative frequencies from each file. Since the semantic and PoS comparisons are inferred from the text words, word differences will be illustrated within semantic and PoS analysis.

3.1 ALICE against Spoken Professional American English transcripts

The Corpus of Spoken Professional American English (CSPA) (Athelstan 2002) includes transcripts of conversation of different types, occurring between 1994 and 1998, covering professional activities broadly tied to academia and politics. The transcripts were recorded during professional meetings. The Wmatrix tables (screenshots) below illustrate the most important differences in semantic, PoS and lexical levels between ALICE chatbot dialogues (file 1) and spoken professional transcripts (file 2).

3.1.1 Semantic comparison between chatbot dialogue and professional American English

Table 1: Semantic comparison between ALICE (01) and CSPA (02)

Sorted by log-likelihood value						
Item	**O1**	**%1**	**O2**	**%2**	**LL**	
E2+	16	1.42	62	0.15	+ 40.90	Liking
Z4	38	3.37	400	0.95	+ 40.05	Discourse bin
Q2.2	37	3.28	449	1.06	+ 32.05	Speech acts
Z1	34	3.01	406	0.96	+ 30.16	Personal names
P1	1	0.09	671	1.59	- 27.63	Education in general
Z8	214	18.97	5485	12.98	+ 26.46	Pronouns etc.
H4	8	0.71	23	0.05	+ 24.20	Residence
X2.2+	19	1.68	173	0.41	+ 23.85	Knowledge
Z5	268	23.76	13314	31.51	- 22.95	Grammatical bin
O4.2+	6	0.53	14	0.03	+ 20.10	Judgement of appearance (pretty etc.)

The semantic comparison in Table 1 shows that the following semantic categories are used more in ALICE transcripts: explicit speech act expressions are highly used within ALICE, an attempt to reinforce the impression that there is a real dialogue; pronouns (e.g. *he*, *she*, *it*, *they*) are used more in ALICE, to pretend personal knowledge and contact; discourse verbs (e.g. *I think*, *you know*, *I agree*) are overused in ALICE, to simulate human trust and opinions during the chat; liking expressions (e.g. *love*, *like*, *enjoy*) are overused in ALICE, to give an impression of human feelings. The only categories used noticeably more in CSPA Professional American English are education terms, hardly surprising given the academic discourse source; and grammatical function words, corresponding to more complex grammar.

3.1.2 Part-of-Speech comparison between chatbot dialogue and professional American English

Table 2: Part-of-Speech comparison between ALICE (01) and CSPA (02)

Sorted by log-likelihood value						
Item	**O1**	**%1**	**O2**	**%2**		**LL**
PPY	80	7.09	503	1.19	+	144.18
VD0	43	3.81	258	0.61	+	80.57
PPIS2	1	0.09	799	1.89	-	34.03
PPIO1	10	0.89	38	0.09	+	25.87
CC	10	0.89	1343	3.18	-	25.68
PPIS1	55	4.88	984	2.33	+	23.02
NP1	44	3.90	744	1.76	+	20.97
NNB	5	0.44	8	0.02	+	19.59
DD1	9	0.80	1030	2.44	-	16.56
CST	6	0.53	813	1.92	-	15.68
UH	14	1.24	181	0.43	+	11.01

Table 2 shows the Part-of-Speech frequency differences between the two sources. Singular first-person pronoun (e.g. *I*), second-person pronoun (e.g. *you*) and proper names (e.g. *Alice*) are used more in ALICE, to mark participant roles more explicitly and hence reinforce the illusion that the conversation really has two participants. Plural personal pronouns (e.g. *we*) were used more in Professional American English, because all samples were extracted from meetings between cooperating professionals, using inclusive language. Coordinating conjunctions (e.g. *and*, *or*) and subordinating conjunctions (e.g. *if*, *because*, *unless*) are used more within Professional American English; these indicate more complex clause and phrase structure, which ALICE avoids because it applies simple pattern matching techniques, and so it has problems trying to handle dependencies between clauses. Professional American English makes less use of interjections, preferring more formal clause structure; another interpretation of this imbalance could be that ALICE makes more use of interjections, as fillers when no good match is found in the pattern database.

3.1.3 Lexical comparison between chatbot dialogue and professional American English

Word-level analysis results shown in table 3 confirm and exemplify the more general Part-of-Speech and semantic class preferences. ALICE transcripts made more use of specific proper names “Alice” (not surprisingly!) and “Emily”; and of “you_know”, where the underscore artificially creates a new single word from two real words. ALICE and human dialogue corpora also made more use of lexical items which correspond to the “marked” PoS and semantic categories above; for example, Alice transcripts included more use of “I”, “you”.

Table 3: Word comparison between ALICE (01) and CSPA (02)

Sorted by log-likelihood value						
Item	**O1**	**%1**	**O2**	**%2**		**LL**
you	72	6.38	496	1.17	+	119.80
Emily	9	0.80	0	0.00	+	65.69
do	44	3.90	370	0.88	+	60.25
you_know	8	0.71	7	0.02	+	38.04
Alice	5	0.44	0	0.00	+	36.50
created	5	0.44	0	0.00	+	36.50
internet	5	0.44	0	0.00	+	36.50
name	6	0.53	2	0.00	+	34.90
we	1	0.09	799	1.89	-	34.03
Dr	4	0.35	0	0.00	+	29.20
chocolate	4	0.35	0	0.00	+	29.20
dance	4	0.35	0	0.00	+	29.20
french	4	0.35	0	0.00	+	29.20
ok	4	0.35	0	0.00	+	29.20
am	6	0.53	5	0.01	+	28.90

The above comparison shows that when ALICE tries to simulate real dialogue, it over-exaggerates use of key lexical, grammatical and semantic features of dialogue. We compared several other DDC human dialogue corpus texts against ALICE transcripts (Abu Shawar and Atwell 2003a); there are genre- or topic-specific differences for each Corpus, but ALICE’s over-exaggerated use of speech act verbs, first-person pronouns, and similar explicit dialogue cues are a recurring result.

4 Training ALICE with conversation transcripts

It took several years for the ALICE Botmaster to accumulate the 50,000 categories in the current public-domain set of AIML files (Wallace 2003). We developed a program to read text from a corpus and convert it to the AIML format. We used this program to generate chatbots speaking other languages including Afrikaans (Abu Shawar and Atwell 2003b) and Arabic (Abu Shawar and Atwell 2004). To retrain ALICE with English corpora, we examined these sources of English dialogue: the dialogue Diversity Corpus (DDC) and the spoken part of the British National Corpus (BNC), and online FAQ (Frequently Asked Questions) websites. Two main goals are achieved using the BNC: the ability of the program to generate more than one million categories extracted from the BNC; and the ability to use the chatbot as a tool to explore and visualise or animate the domain-specific English dialogue.

4.1 Problems with using BNC as chatbot training data

The British National Corpus (BNC) (Aston and Burnard 1998) is a collection of text samples amounting to over 100 million words, extracted from 4,124 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. In order to generate versions of ALICE which talk like humans, we decided to retrain ALICE using the BNC spoken transcripts comprising 10 million words.

Sample 1.a illustrates the long monolog problem

```
<u who=F72PS002>
<s n="29"><w PNP>You <w VDB>do<c PUN>?
<s n="30"><w AV0>Well <w PNP>you <w VBB>are <w AV0>very <w
AJ0>fortunate <w NN0>people<c PUN>.
<s n="31"><w CJC>But <w PNI>none <w PRF>of <w PNP>you <w VM0>will
<w VVI>know <w DPS>my <w NN1>friend <w AV0>over here <w
DTQ>whose <w NN1>name <w VBZ>is
<w NP0>Donald<c PUN>. </u>
```

Sample 1.b illustrates the overlapping problem

```
<u who=w0014>
<s n=00011><w AJ0>Poor <w AJ0>old <w NP0>Luxembourg'<w VBZ>s
<w AJ0-VVN>beaten<c PUN>.
<s n=00012><w PNP>You <w PNP>you<w VHB>'ve <w PNP>you<w VHB>'ve
```

```
<w AV0>absolutely <w AV0>just<w VVN>gone <w AV0>straight
<ptr target=P1> <w PRP>over <w PNP>it <ptr target=P2> </u>
<u who=w0001>
<s n=00013><ptr target=P1> <w PNP>I <w VHB>haven<w XX0>'t<c PUN>.
<ptr target=P2/> </u>
<u who=w0014>
<s n=00014><w CJC>and <w VVN>forgotten <w AT0>the <w AJ0>poor
<w AJ0>little<w NN1>country<c PUN>. </u>
```

Figure 1: Samples of BNC corpus illustrating some problems

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 (anonymous, e.g. 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. Some example utterances are shown in Figure 1.

The translation process from BNC format to AIML is not as simple as it might seem to be on the surface. A range of problems emerged during the translation process:

- More than two speakers and long monologues as illustrated in sample 1.a
- Unclear sections and overlapping as illustrated in sample 1.b
- Other problems including: extra-linguistic annotations, abbreviations, and using character entity references.

4.2 Using the BNC word-frequency list

The basic aim of the program is to build categories form the existing turns. The simplest way is by considering each turn to be a pattern and the next one to be a template in an atomic category. However, the user inputs can not be restricted to these atomic categories; to extend the possibility of finding a match, a "most significant word" approach was adopted. The most significant word was extracted from each utterance, the word in the utterance with the highest "information content", the word that is most specific to this utterance compared to other utterances in the corpus. This should be the word that has the lowest frequency in the rest of the corpus. We choose the most significant approach to generate the default categories, because usually in human dialogues the intent of the speakers is hiding in the least-frequent, highest-information word. To obtain the least-frequent word from the BNC we utilised, the unlemmatised list implemented by Kilgarriff (1996). The unlemmatised list counts the frequency for each verbal instance separately. There are six forms of the unlemmatised list: sorted alphabetically, or by frequency, complete lists, a smaller file containing only those items occurring over five times, and compressed and uncompressed versions of all available lists. The unlemmatised list format compromises four fields separated by spaces, starting with the frequency, the word, the POS, and finally the number of files the word occurs in, as shown Figure 2.

```
6187267    the   at0   4120
2941444    of    prf   4108
2682863    and   cjc   4120
2126369    a     at0   4113
1812609    in    prp   4109
```

Figure 2: Sample of the (Kilgarriff 1996) BNC word-frequency list

Since the BNC spoken transcripts are annotated with Part-of-Speech tags, we used these tags to extract the meaningful words to be used as the first word or least frequent words. Questions, prepositions, and pronouns are not considered. This modification improves the matching process and we recorded better user satisfaction than before.

5 A chatbot as a tool to visualize or animate corpus texts

One possible use of a corpus-trained chatbot is as a novel corpus visualization tool; Abu Shawar and Atwell (2004b) suggested:

> … the classical way of viewing a data set is using a 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. … We introduce a new idea to visualize a dialogue corpus using a chatbot interface tool. … we use specific domains of the BNC spoken files to retrain ALICE, and visualise the data contents of these domains via chatting…. 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… .

Information Visualisation techniques are usually applied to large numerical datasets, but linguists do have some tools for summarising and abstracting overview information from textual datasets. For example, Wmatrix presents summary statistical information comparing two corpora; although strictly speaking the output is not a "picture" of the corpora, it can still be seen as a visualisation: "In general, the notion 'visualization' refers to the process of mapping (non-visual) data to a suitable representation that simplifies the data interpretation by the user" (MoVi 1996). In contrast to Wmatrix, the chatbot can give qualitative impressions about a corpus. The chatbot could be used as a tool to visualise the vocabulary and typical turn-taking within the context, in addition to viewing the type of language used. Perhaps a better description for use of a corpus-trained chatbot to explore a corpus is "animate": "to give life to or cause to come alive" (*Collins English dictionary*). A range of different chatbot systems were generated using the program that converts the BNC to AIML format; this enabled us to animate and explore the language of various BNC sample texts. 1,153,129 categories were created in total, divided in files representing a range of domains as illustrated in Table 4.

Table 4: Sample of AIML chatbot models learnt from BNC domains

Chatbot	**Atomic**		**Default**		**Total**
	Categories	**File size (KB)**	**Categories**	**File size (KB)**	**Categories**
Sport	3283	974	7913	1,820	11196
World affairs	3120	983	8756	1,886	11876
Travel	640	314	1636	575	2276
Media	1061	491	3126	1,210	4187
Food	503	93	1125	168	1628

Human: Hi, do you like sport?
Chatbot: Hi there. No I'm not much into sports. I like solo sports like rollerblading and skiing though.
Human: Do you know anything about football?
Chatbot: 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?
Chatbot: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?
Chatbot: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.
Chatbot: 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.
Chatbot: 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?
Chatbot: 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?
Chatbot:mm

Chatting 1: Sample of chat with BNC trained Sports chatbot

For example, chatting 1 (from Abu Shawar and Atwell 2004b) illustrates chatting with the BNC-Sports chatbot. Notice that the ALICE pattern-matching approach allows ill-formed or even ungrammatical input to be matched partially; for example, the second input from the user was presumably meant to be "*Do you know anything about football?*" but the user inadvertently left out the verb. The chatbot replies may also be odd or ungrammatical, as they directly reflect the language of the training corpus. This chatting example also shows there may be logical inconsistencies in the chatbot replies; for example, early on the chatbot says: "no I'm not much into sport", but later it enthuses about football. This is because the chatbot was trained with a range of BNC texts, not just a single speaker, so it models an amalgam of all the conversations about sport; also, the ALICE architecture does not maintain a “belief model”, so it has no way of avoiding logical inconsistencies across the chatting session. However, it is possible to visualize or animate other aspects of language use: for example, longer turns when the chatbot (speaker) gets enthused; use of domain specific collocations or idioms in context (eg “its’ poetry in motion”); gender variation in language (the speaker in the training text was obviously a man; a chatbot trained with transcripts of women talking about sport would respond differently).

Table 5: Sample of AIML chatbot models learnt from BNC London teenager and “loudmouth” transcripts

Teenager version	**Atomic**		**Default**		**Total**
	Categories	**File size (KB)**	**Categories**	**File size (KB)**	**Categories**
Michael	7021	1,044	14914	1,905	21935
Jessica	4261	787	10069	1,554	14330
Peter	5912	939	12878	1,771	18790
Robin	4726	673	14580	1,774	19306
Samantha	6058	788	18283	1,695	24341
Loudmouth1	8114	1,301	17827	2,368	25941
Loudmouth2	8383	1,217	16971	2,130	25354

Human: I never give cigarettes
Peter: Did you give that one back to Dan and Honey?
Human: You would not give me a cigarette anyway
Peter: In any case, Who else would not give you a cigarette?
Human: It depends
Peter: Yeah
Human: Did your mum know that you smoke cigarettes
Peter: About a month ago

Chatting 2: Sample of chat with BNC London teenager Peter

A significant subset of BNC spoken conversation transcripts are of London teenager chat collected by the COLT project. To animate these, we built several London-teenager chatbots trained on specific transcripts. We also posted a query to ICAME's CORPORA email discussion list, asking for suggestions for "interesting" training texts in the BNC; replies suggested two texts to train two chatbots, Loudmouth1 and Loudmouth2, to illustrate loudmouth vocabulary and dialogue. Table 5 shows statistics for some of the AIML models extracted from BNC London teenager and loudmouth files. Chatting 2 illustrates dialogue between a user and our Peter chatbot.

Another sort of English "dialogue" text is FAQs: Frequently Asked Questions websites. Corpus Linguists have shown increasing interest in possibilities of using the Web as corpus source, so we decided to try using FAQs as pseudo-dialogue to train chatbots. In these versions the chatbot expects user input similar to FAQ questions, and replies with FAQ answers, or with lists of links to possible FAQ answers. FAQs have clear turn-taking structure, and are widely used in computing, so are a good source of training corpus data.

Two example English FAQ chatbots were generated: Askme (2004) which is trained using the FAQ of the School of Computing (SoC) at the University of Leeds; and a Python advisor (2004), trained on the public-domain Python programming language FAQ website. The Python FAQ website has been translated into other languages, so we were able to generate a Spanish-speaking Python advisor. Chatting 3 illustrates the questions and answers generated with these FAQ chatbots.

---

a. Using the FAQ of the School of Computing at the University of Leeds

Human: How to remove a job from the printer?
Ask me:

If you want to remove a job from the printer queue then find out the job number using the lpq command, and in a UNIX shell type:

% **lprm -P** *printername job_no*

You can only remove jobs that you have sent to the printer yourself - apart from Support that is.

For more information look at:
How do I remove a job from the printer queue? (Printing)

b. Using the FAQ of the Python programming language website

Human: what is a class ?
English_Python_FAQ:

A class is the particular object type created by executing a class statement. Class objects are used as templates to create instance objects, which embody both the data (attributes) and code (methods) specific to a datatype.

A class can be based on one or more other classes, called its base class(es). It then inherits the attributes and methods of its base classes. This allows an object model to be successively refined by inheritance. You might have a generic Mailbox class that provides basic accessor methods for a mailbox, and subclasses such as MboxMailbox, MaildirMailbox, OutlookMailbox that handle various specific mailbox formats.

Chatting 3: Example questions with answers generated by the FAQ chatbot versions

A Frequently Asked Questions document or FAQ is a structured database; most of the “data-cleaning” problems found with dialogue corpora (such as overlapping, and more than two speakers) are not found in the FAQs. Moreover almost all HTML tags embedded in the FAQ file are recognised by the ALICE interpreter, because the AIML definition allows HTML tags to be embedded within templates. The questions and answers were extracted directly from the HTML files of the FAQ.

6 Conclusions

A chatbot is a machine conversation system, which enables users to interact with it using natural language. Existing chatbot transcripts are different from dialogue corpus texts, as they show exaggerated use of speech-act verbs, first and second singular person pronouns, and similar explicit linguistic cues of dialogue. Most chatbots require linguistic knowledge to be hand coded, and restricted to the language in which it is written. We developed a Java program that uses a corpus-based approach to train ALICE. In this paper the British National Corpus spoken transcripts were used to retrain ALICE. We were able to develop the most significant word approach as a learning technique, and were successful in learning a range of illustrative chatbots from the BNC corpus. Two goals were achieved from the automation process: using the BNC corpus to automatically generate the largest language processing model ever, containing 1,153,129 categories or dialogue-rules; and using chatbots trained on specific subsets of the BNC to animate and explore the type of the language used within a specific domain. In contrast to the Wmatrix tool which gives a quantitative view about the syntax, semantics and PoS used within the corpus, the chatbot system gives a qualitative illustration or animation of the language of a specific domain or speaker type.

Future research includes finding useful applications. The ALICE-style chatbot could be used as an English conversation practice tool; for example Speak2Me.net (Speak2Me 2004) is a web-based version of ALICE aimed at Chinese learners of English, allowing them to practise chatting to a well-spoken young lady, a virtual British English native speaker. Our chatbot-training technique could be used in producing chatbots aimed at learners with other first languages, e.g. German or Italian (Atwell et al. 2003), or chatbots specialising in English for specific purposes, by retraining ALICE with BNC corpus (or other ESP corpus texts) on specific domains.

References

Abu Shawar, Bayan and Eric Atwell. 2002. A comparison between ALICE and Elizabeth chatbot systems. *School of Computing research report 2002.19*, University of Leeds.

Abu Shawar, Bayan and Eric Atwell. 2003a. Using dialogue corpora to retrain a chatbot system. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Proceedings of the Corpus Linguistics 2003 Conference (CL2003)*, 681-690. Lancaster: Lancaster University.

Abu Shawar, Bayan and Eric Atwell. 2003b. Using the Corpus of Spoken Afrikaans to generate an Afrikaans chatbot. *Southern African Linguistics and Applied Language Studies* 21: 283-294.

Abu Shawar, Bayan and Eric Atwell. 2004a. Accessing an information system by chatting. In F. Meziane and E. Metais (eds.). *Natural Language Processing and Information Systems, 9th International Conference on Applications of Natural Languages to Information Systems. Proceedings (*NLDB'04), 407-412. Berlin: Springer.

Abu Shawar, Bayan and Eric Atwell. 2004b. A chatbot as a novel corpus visualization tool. In M. Teresa Lino, M. Francisca Xavier, F. Ferreira, R. Costa, R. Silva, C. Pereira, F. Caevalho, M. Lopes, M. Catarino and S. Barros (eds.). *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC'04)* IV, 2057-2060. Paris: ELRA.

ALICE. 2002. *A.L.I.C.E AI Foundation*, http://www.alicebot.org/

Aston, Guy and Lou Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atwell, Eric, Peter Howarth and Clive Souter. 2003. The ISLE corpus: Italian and German spoken learner's English. *ICAME Journal* 27: 5-18.

Burnard, Lou and Tony Dodd. 2003. Xara: An XML aware tool for corpus searching. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.). *Proceedings of the Corpus Linguistics 2003 Conference (CL2003)*, 142-144. Lancaster: Lancaster University.

Colby, Kenneth. 1973. Simulation of belief systems. In R. Schank and K. Colby (eds.). *Computer models of thought and language*, 251-286. San Francisco: Freeman.

Colby, Kenneth. 1999. Comments on human-computer conversation. In Y. Wilks (ed.). *Machine conversations*, 5-8. Boston, Dordrecht and London: Kluwer.

Cunningham, Hamish. 1999. A definition and short history of Language Engineering. *Journal of Natural Language Engineering* 5.1: 1-16.

Güzeldere, Güven and Stefano Franchi. 1995. Dialogues with colourful personalities of early ai. *Stanford Electronic Humanities Review* 4.2, http://www.stanford.edu/group/SHR/4-2/text/toc.html

Kilgarriff, Adam. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography* 10.2: 135-155.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In J. Svartvik (ed.). *Directions in corpus linguistics*, 105-122. Berlin: Mouton de Gruyter.

Leech, Geoffrey, Roger Garside and Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In P*roceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, 622-628. Kyoto, Japan.

Loebner, Hugh. 2003. *Home Page of the Loebner Prize-The First Turing Test*, http://www.loebner.net/Prizef/loebner-prize.html

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

MIA. 2004. http://www.aitools.org/livebots/

MoVi. 1996. *MoVi Glossary*, http://wwwicg.informatik.uni-rostock.de/Projekte/MoVi/mirror/a_notions.html

MUMMI. 2004. *A Concept for chatbot: "Sanelma" building engaging relationship between the work of art and the exhibition visitor,* http://www.mlab.uiah.fi/mummi/sanelma/

Rayson, Paul. 2002. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University, UK.

Sinclair, John McH. 1991. *Corpus, concordance, collocation*. Oxford: OUP.

Speak2Me. 2004 www.speak2me.net

Spielberg Stephen. 2000. *Random conversation with a chatbot*, http://www.aimovie.warnerbros.com/

Wallace, Richard. 2003. *The elements of AIML style.* ALICE AI Foundation.

Weizenbaum, Joseph. 1966. ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM* 10.8: 36-45.

Wilks, Yorick. 1999. Preface. In Y.Wilks (ed.). *Machine conversations,* vii-x. Boston, Dordrecht and London: Kluwer.