

Evaluation of Chatbot Information System

Bayna Abu shawar, Eric Atwell

*School of Computing, University of Leeds, Leeds LS2 9JT
bshawar@comp.leeds.ac.uk*

*School of Computing, University of Leeds, Leeds LS2 9JT
eric@comp.leeds.ac.uk*

ABSTRACT – *A chatbot is a software system, which can interact or “chat” with a human user in natural language such as English. For the annual Loebner Prize contest, rival chatbots have been assessed in terms of ability to fool a judge in a restricted chat session. We are investigating methods to train and adapt a chatbot to a specific user’s language use or application, via a user-supplied training corpus. Our evaluation takes account linguistically-motivated comparison of human dialogue and chatbot transcripts. We also advocate open-ended trials by real users, such as an example Afrikaans chatbot for Afrikaans-speaking researchers and students in South Africa. This is evaluated in terms of “glass box” dialogue efficiency metrics, “black box” dialogue quality metrics and user satisfaction feedback. Our general conclusion is that evaluation should be adapted to the application and to user needs.*

KEY WORDS: *a chatbot, Loebner Prize contest, Wmatrix, evaluation*

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 99). Human machine conversation as a technology integrates different areas where the core is the 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 such as ELIZA (Weizenbaum 66, 67), and ALICE. Chatbots have been used in different domains such as: customer service, education, web site help, and for fun.

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 03) 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 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”. Different mechanisms are used to evaluate Spoken Dialogue Systems, ranging from glass box evaluation that evaluates individual components, to black box evaluation that evaluates the system as a whole.

The main goal of conversational machines is to mimic human conversations. Section 2 illustrates the Loebner Prize contest and Alice architecture. In this paper the evaluation process will be tackled in two directions. First, section 3 compares human-to-human dialogues with human to machines dialogues according to significant differences in lexis or vocabulary, grammar, and semantics. Secondly, section 4 uses the black box mechanism to evaluate user satisfaction with our example Afrikaans version of ALICE. The conclusion is presented in section 5.

2. BACKGROUND

2.1. The Loebner Prize Competition

The story began with the “imitation game” and the question “Can machines think?” which were presented by Alan Turing (Turing 50). The imitation game has a human observer who tries to guess the sex of two players, one of which is a man and the other is a woman, but while screened from being able to tell which is which by voice, or appearance. Turing suggested putting a machine in the place of one of the humans and essentially playing the same game. If the observer cannot tell which is the machine and which is the human, this can be taken as strong evidence that the machine can think.

Turing’s proposal provided the inspiration for the Loebner Prize competition, which was an attempt to implement the Turing test. The first contest organized by Dr. Robert Epstein was held on 1991. Ten agents were used, 6 were computer programs. Ten judges would converse with the agents for fifteen minutes and rank the terminals in order from the apparently least human to most human. The computer with the highest median rank wins that year’s prize.

However there are sceptics who doubt the effectiveness of the Turing Test and/or the Loebner Competition. Searle (1980) objected to Turing’s idea that the machine is thinking if it can fool users into believing that they are speaking to a human. Other sceptics include Shieber (1994), claimed the reason that Turing chose natural language as the behavioural definition of human intelligence is “exactly its open-ended, freewheeling nature”, which was lost when the topic was restricted during the Loebner Prize.

Loebner in his responses to these arguments believed that unrestricted test is simpler, less expensive and the best way to conduct the Turing Test. Loebner presented three goals when constructing the Loebner Prize (Loebner 03). These are: increasing the public understanding of AI, performing a social experiment, and such contest was the first time the Turing Test had ever been formally tried.

ALICE¹, the Artificial Linguistics Internet Computer Entity has won the Loebner Prize twice, in 2000 and 2001. Dr. Richard Wallace implemented ALICE in 1995, and gradually extended its rule-set, to cope with open-ended conversation. A depth first search algorithm with backtracking is used to find the best match for specific user input.

Whether or not the Loebner Prize advances the field of Artificial Intelligence, it does make us aware of how little our understanding of conversation lies in what is said. The annual Loebner Prize contest encourages researchers to develop chatbots that can pass the competition. However this has led developers to focus on ways to meet the 10-minute challenge, rather than on how to build practical, useful information systems.

2.2. ALICE System Architecture

ALICE stores knowledge about English conversation patterns in AIML files. AIML, or Artificial Intelligence Mark-up Language, is a derivative of Extensible Mark-up Language (XML). It was developed by the Alicebot free software community during 1995-2000 to enable people to input dialogue pattern knowledge into chatbots based on the ALICE free 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 unit of knowledge in AIML. 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. The idea of the pattern matching technique is based on finding the best, longest, pattern match.

2.2.1. Types of ALICE/AIML Categories

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

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

```
<category><pattern>10 Dollars</pattern>
<template>Wow, that is cheap. </template></category>
```

In the above category, if the user inputs ‘10 dollars’, then ALICE answers ‘WOW, that is cheap’.

– Default categories are those with patterns having wildcard symbols `*` or `_`. The wildcard symbols match any input but they differ in their alphabetical order. Assuming the previous input 10 Dollars, if the robot does not find the previous category with an atomic pattern, then it will try to find a category with a default pattern such as:

```
<category><pattern>10 *</pattern><template>It is ten.</template> </category>
```

So ALICE answers ‘It is ten’.

– Recursive categories are those with templates having `<sr>` 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 as the following example:

```
<category><pattern>HIYA</pattern><template><srail>Hello</srail></template>
</category>
```

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

2.2.2. ALICE/AIML Pattern Matching Technique

The AIML interpreter tries to match word by word to obtain the longest pattern match, as this is normally the best one. This behaviour can be described in terms of the Graphmaster set of files and directories, which has a set of nodes called nodemappers and branches representing the first words of all patterns and wildcard symbols. Assume the user input starts with word X and the root of this tree structure is a folder of the file system that contains all patterns and templates; the pattern matching algorithm uses depth first search techniques:

If the folder has a subfolder starting with underscore then turn to, “_”, scan through it to match all words suffixed X, if no match then:

Go back to folder, try to find a subfolder starts with word X, if so turn to “X/”, scan for matching the tail of X, if no match then:

Go back to the folder, try to find a subfolder start with star notation, if so, turn to “*/”, try all remaining suffixes of input following “X” to see if one match. If no match was found, change directory back to the parent of this folder, and put “X” back on the head of the input.

When a match is found, the process stops, and the template that belongs to that category is processed by the interpreter to construct the output.

3. HUMAN TO HUMAN VERSUS HUMAN TO CHATBOT DIALOGUES

In this section we will compare a dialogue transcript generated via chatting with ALICE, and real conversations extracted from different dialogue corpora. The comparison will illustrate the strength or weakness of ALICE as a human simulation, according to linguistic features: lexical, Part-of-Speech, and semantic differences. The Wmatrix tool (Rayson 02) 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 02), and human-to-

computer dialogues extracted from chats with ALICE on the AI movie website (Spielberg 00). Four different corpora in different fields and sizes were investigated; the DDC sub corpora and ALICE transcript are not equal in size, so we will look to the ratio value of 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 02) 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.

3.1.1. Part-of-Speech comparison

Part-of-Speech comparison shows that the 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 more used within Professional American English, these indicate more complex clause and phrase structure, which ALICE avoids because it applies simple pattern matching techniques, and cannot 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.

3.1.2. Semantic Comparison

Semantic comparisons 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 more used 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.

3.1.3. Lexical comparison

Word-level analysis results shown in screenshot 1 (where O1 represents ALICE dialogue and O2 represents spoken professional transcripts) 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”.

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
name	6	0.53	2	0.00 +	34.90
we	1	0.09	799	1.89 -	34.03
am	6	0.53	5	0.01 +	28.90

Screenshot 1. Word comparison

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 compare several other human dialogue corpus texts against ALICE transcripts; 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. ALTERNATIVES TO THE LOEBNER PRIZE USER EVALUATION APPROACH

Instead of chatting for just 10 minutes as suggested by the Loebner Prize, we advocate alternative evaluation methods more attuned to and appropriate to practical information systems applications. We are investigating methods to train and adapt a chatbot to a specific user’s language use or application, via a user-supplied training corpus (Abu Shawar et al., 02, 03a). Our evaluation takes account of open-ended trials by real users, rather than artificial 10-minute trials. One example is a chatbot for Afrikaans-speaking researchers and students in South Africa (Abu Shawar et al., 03b).

4.1. Afrikaans chatbot version

We have adapted the ALICE/AIML chatbot architecture to be retrained from a dialogue corpus to generate a new version of ALICE in a different language style (eg Professional American English), or even a completely different language. We were supplied with a training corpus of Afrikaans dialogue transcripts, the Korpus Gesproke Afrikaans (van 02), and used this to develop Afrikaans-speaking and bilingual Afrikaans-English chatbots. Our first attempt was based only on literal pattern matching against corpus utterances: where the atomic categories are used to match the user input with its corresponding exact pattern. However this method was not satisfying for users, since it requires the input to be exactly the same as the pattern; too often, no pattern was found to exactly match user input, so the response was just a default “ja” (“yes”). In order to generate a more interesting answer for more of the user input, we adopted two generalisation approaches or strategies:

- First word approach: this is based on the generalisation that the first word of an utterance may be a good clue to an appropriate response; if we can not match the whole input utterance, then at least we can try matching just the first word. For each atomic pattern, we generate a default version that holds the first word followed by star to match any text, and then associated it with the same atomic template.

- The most significant word approach: this is loosely based on information theory. Instead of assuming the first word of an utterance is most “significant”, we look for 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 tokenised the whole corpus into word-tokens, to produce a list of all words, along with the frequency of each word. Sorting this list in ascending order yields the overall least frequent word list. Then we tokenised each pattern and compared it with the corpus word-frequency list to get the least frequent word in this pattern. Four default categories are added which hold this least-frequent word in different position of the pattern: start, middle, last or just the least frequent word alone. We choose the least frequent 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.

4.2 Evaluation of the Afrikaans chatbots

We developed two versions of the ALICE that speaks Afrikaans language, Afrikaana that speaks only Afrikaans and AVRA that speaks English and Afrikaans; this was inspired by our observation that the Korpus Gesproke Afrikaans actually includes some English, as Afrikaans speakers are generally bilingual and “code-switch” comfortably. We mounted prototypes of the chatbots on websites using Pandorobot service (Pandorobot 03), and encouraged open-ended testing and

feedback from remote users in South Africa; this allowed us to refine the system more effectively. We adopted three evaluation metrics:

- Dialogue efficiency in terms of matching type.
- Dialogue quality metrics based on response type.
- Users satisfaction assessment based on an open-ended request for feedback.

4.2.1. Dialogue efficiency metrics

We measured the efficiency of 4 sample dialogues in terms of atomic match, first word match, most significant match, and no match. We wanted to measure the efficiency of the adopted learning mechanisms to see if they increase the ability to find answers to general user input. Table 1 shows the frequency of each type in each dialogue generated between the user and the Afrikaans chatbot; in Figure 1, these absolute frequencies are normalised to relative probabilities.

Matching Type	Dialogue 1	Dialogue 2	Dialogue 3	Dialogue 4
Atomic	1	3	6	3
First word	9	15	23	4
Most significant	13	2	19	9
No match	0	1	3	1
Number of turns	23	21	51	17

Table 1. *Response Type Frequencies*

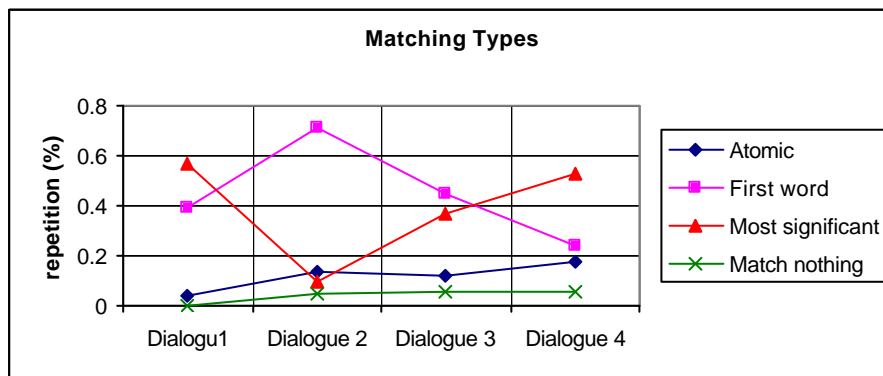


Figure 1. *Dialogue efficiency: Response Type Relative Frequencies*

This approach to evaluation via dialogue efficiency metrics illustrates that the first word and the most significant approach increase the ability to generate answers to users and let the conversation continue.

4.2.2. Dialogue quality metrics

In order to measure the quality of each response, we wanted to classify responses according to an independent human evaluation of “reasonableness”: reasonable reply, weird but understandable, or nonsensical reply. We gave the transcript to an Afrikaans-speaking teacher and asked her to mark each response according to these classes. Table 2 shows the number of turns in each dialogue and the frequencies of each response type. Figure 2 shows the frequencies normalised to relative probabilities of each of the three categories for each sample dialogue. For this evaluator, it seems that “nonsensical” responses are more likely than reasonable or understandable but weird answers.

Response Type	Dialogue 1	Dialogue 2	Dialogue 3	Dialogue 4
Number of turns	23	21	51	17
Reasonable	2	4	5	5
Weird	19	3	7	1
Nonsensical	2	14	39	11

Table 2. *Response types frequencies*

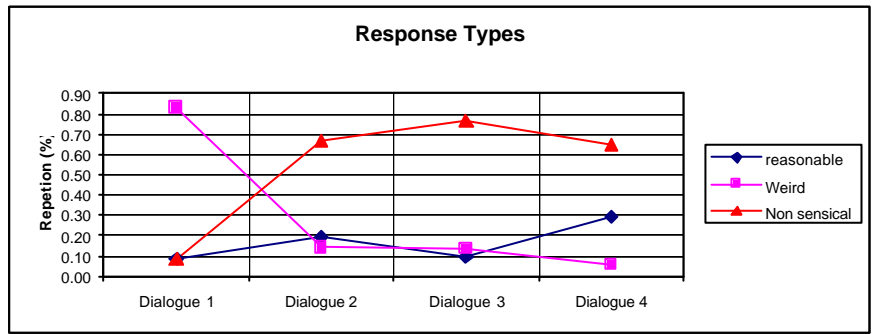


Figure 2. *The quality of the Dialogue: Response type relative probabilities*

4.2.3. User satisfaction

The first prototypes were based only on literal pattern matching against corpus utterances: we had not implemented the first word approach and least-frequent word

approach to add “wildcard” default categories. Our Afrikaans-speaking evaluators found these first prototypes disappointing and frustrating: it turned out that few of their attempts at conversation found exact matches in the training corpus, so Afrikaana replied with a default “ja” most of the time. However, expanding the AIML pattern matching using the first-word and least-frequent-word approaches yielded more favourable feedback: our informants found the conversations less repetitive and more interesting. We measure user satisfaction based on this kind of informal user feedback.

5. CONCLUSIONS

The Loebner Prize Competition has been used to evaluate the ability of chatbots to fool people that they are speaking to humans. Comparing the dialogues generated from ALICE, which won the Loebner Prize with real human dialogues, shows that ALICE tries to use explicit dialogue-act linguistic expressions more than usual to reinforce the impression that that users are speaking to human.

Expanding the AIML pattern matching using the first-word and least-frequent-word approaches yielded more favourable feedback: our informants found the conversations less repetitive and more interesting, although responses were sometimes weird and apparently irrelevant or nonsensical in context. The reasons behind most of the users’ feedback can be related to three issues. Firstly the dialogue corpus context does not cover a wide range of domains, so Afrikaana can only “talk about” the domain of the training corpus. Secondly, the repeated approach that we used to solve the problem of determining the pattern and the template in case of more than two speakers may lead to incoherent transcripts. Thirdly, our machine-learned models have not included linguistic analysis markup, such as grammatical, semantic or dialogue-act annotations (Atwell et al 00), as ALICE/AIML makes no use of such linguistic knowledge in generating conversation responses.

Our general conclusion is that we should not adopt an evaluation methodology just because a standard has been established, such as the Loebner Prize evaluation methodology adopted by most chatbot developers. Instead, evaluation should be adapted to the application and to user needs. If the chatbot is meant to be adapted to provide a specific service for users, then the best evaluation is based on whether it achieves that service or task

6. REFERENCES

- [Abu et al. 02] Abu Shawar B., Atwell E., “A comparison between ALICE and Elizabeth chatbot systems.”, School of Computing research report 2002.19, University of Leeds.

- [Abu et al. 03a] Abu Shawar B., Atwell E., "Using dialogue corpora to retrain a chatbot system". In *Proceedings of the Corpus Linguistics 2003 conference*, Lancaster University, UK, p. 681-690.
- [Abu et al. 03b] Abu Shawar B., Atwell E., "Using the Corpus of Spoken Afrikaans to generate an Afrikaans chatbot". To appear in *SALALS Journal: Southern African Linguistics and Applied Language Studies*.
- [Athelstan 02] Athelstan, "Corpus of Spoken Professional American-English": description, <http://www.athel.com/corpdes.html>
- [Atwell et al. 00] Atwell E., Demetriou G., Hughes J., Schiffrin A., Souter C., Wilcock S., "A comparative evaluation of modern English corpus grammatical annotation schemes." *ICAME Journal* 24: p. 7-23.
- [Colby 99] Colby K., "Comments on Human-computer conversation. In *Machine conversations*, Yorick Wilks (Eds.), Kluwer, Boston/Dordrecht/London, p. 5-8.
- [Cunningham 99] Cunningham H., "A definition and short history of Language Engineering". *Natural Language Engineering* 5(1), p.1-16.
- [Loebner 03] Loebner H., "Home Page of the Loebner Prize-The First Turing Test", [Online], <http://www.loebner.net/Prizef/loebner-prize.html>
- [Mann 02] Mann W., "Dialog Diversity Corpus" [Online] <http://www-rcf.usc.edu/~billmann/diversity/DDivers-site.htm>
- [Pandorabot 03] Pandorabot, "Pandorabot chatbot hosting service", [online], <http://www.pandorabots.com/pandora>
- [Rayson 02] Rayson P., "Matrix: a statistical method and software tool for linguistic analysis through corpus comparison", Ph.D. thesis, Lancaster University.
- [Searle 80] Searle J., "Minds, Brains, and Programs", *The behavioural and Brain Sciences*, vol 3, No.3, p. 417-457.
- [Shieber 94] Shieber S., "Lessons from a Restricted Turing Test", *Communications of the Association for Computing Machinery*, Vol 37, No. 6, p. 70-78
- [Spielberg 00] Spielberg S., "Random conversation with a chatbot", [Online], <http://aimovie.warnerbros.com/>.
- [Turing 50] Turing A., "Computing Machinery and intelligence". *Mind* 59, 236, p. 433-460.
- [Van 02] Van Rooy B., "Transkripsiehandleiding van die Korpus Gesproke Afrikaans" [Transcription Manual of the Corpus Spoken Afrikaans.] Potchefstroom: Potchefstroom University.
- [Weizenbaum 66] Weizenbaum J., "ELIZA-A computer program for the study of natural language communication between man and machine", *Communications of the ACM*, Vol. 10, No. 8, p. 36-45.
- [Weizenbaum 67] Weizenbaum J., "Contextual understanding by computers", *Communications of the ACM*, Vol. 10, No. 8, p. 474-480.