

Machine learning from dialogue corpora to generate chatbots

Bayan Abu Shawar (bshawar@comp.leeds.ac.uk)

and Eric Atwell (eric@comp.leeds.ac.uk)

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

Abstract

A chatbot is a conversational agent that interacts with users turn by turn using natural language. Different chatbots or human-computer dialogue systems have been developed using spoken or text communication and have been applied in different domains such as: linguistic research, language education, customer service, web site help, and for fun. However, these chatbots have so far been restricted to the linguistic knowledge that is "hand-coded" in their files. For example, to port an English-speaking chatbot to converse in another language, an expert in that language must laboriously hand-translate the English linguistic pattern-matching rules into the target language.

International research in NLP is dominated by work on English. NLP techniques and systems can be ported to other natural languages, but this is generally a labour-intensive task, requiring scarce computational and linguistic expertise; hence minority languages are poorly represented in NLP technology. We present an automated approach to porting an NLP technology, the AIML-based chatbot, to new languages, by using a corpus in the target language to retrain the chatbot. We have successfully automated production of chatbots talking different varieties of English, as well as French, and Afrikaans; and are developing further demonstrators in Spanish and Arabic.

1. Introduction

Human machine conversation is a new technology integrating different areas where the core is the language, and the computational methodologies, which aim to facilitate communication between users and computers via natural language. A related term to machine conversation is the chatbot, which is a conversational agent that interacts with users turn by turn using natural language; chatbots have been applied in different domains such as: linguistic research, language education, customer service, web site help, and for fun. ALICE (ALICE 2002, Abu Shawar and Atwell 2002) 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.

Since the primary goal of chatbots is to mimic real human conversations, we developed a java program that learns from dialogue corpora to generate AIML files in order to modify ALICE to behave like the corpus. Chatbots have so far been restricted to the linguistic knowledge that is "hand-coded" in their files. For example, to port an English-speaking chatbot to converse in another language, an expert in that language must laboriously hand-translate the English linguistic pattern-matching rules into the target language. In this paper we present the software implementation and the problems, which arose in learning from a dialogue corpus.

2. Software Implementation: Version 1

We developed a java program that converts the readable text (corpus) to the chatbot language model format. Two versions of the program were generated. The first version is based on simple pattern template category, so the first turn of the speech is the pattern to be matched with the user input, and the second is the template that holds the robot answer. This version was tested using the English-language Dialogue Diversity Corpus (DDC) (Mann 2002), to investigate the problems of utilising

dialogue corpora. The dialogue corpora contain linguistic annotation that appears during the spoken conversation such as overlapping, and using some linguistic fillers. To handle the linguistic annotations and fillers, the program is composed of four phases as follows:

1. Phase One: Read the dialogue text from the corpus and insert it in a vector.
2. Phase Two: Text reprocessing modules, where all linguistic annotations such as overlapping, fillers and other linguistic annotations are filtered.
3. Phase Three: converter module, where the pre-processed text is passed to the converter to consider the first turn as a pattern and the second as a template. Removing all punctuation from the patterns and converting it to upper case is done during this phase.
4. Phase Four: Copy these atomic categories in an AIML file.

In the next section we will present problems, which arose during the extraction process from the DDC.

3. Problems in the Dialog Diversity Corpus

The DDC is a collection of links to different dialogue corpora in different fields. These annotated texts are transcribed from recorded dialogues between more than two speakers.

MICAS Corpus: Michigan Corpus of Academic Spoken English (MICASE 2002), a collection of transcripts of academic speech events recorded at the university of Michigan. *For example*:

Astronomy transcript:

S1: circumpolar stars. So if I keep my pointer there, [S2: oh] <ROTATES CEILING> everything else moves and we all get sick. <SS LAUGH> and we go backwards in time. And that's even more fun.

S2: make it go really really fast.

<SS LAUGH>

S1: Okay so that's how the sky is going to move, a couple of other things that we can do in here, um, this is a presentation of, the, grid, that we use to divide the sky, so these lines that run, north south what do we call those?

S3: declination

The Astronomy transcript was made using a digital audio tape recorder with two microphones. Problems illustrated are: long monologs, overlapping denoted by angle brackets, more than two speakers and extra annotations recorded actions such as <SS LAUGH>.

CIRCLE corpus [6] (Center for interdisciplinary research on constructive learning environments), is a collection of transcripts holding different tutorial sessions, on topics such as physics, algebra and geometry. For example:

Algebra transcript:

TUTOR [Opening remarks and asks student to read out aloud and begin]

STUD [Reads problem] Mike starts a job at McDonald's that will pay him 5 dollars an hour, Mike gets dropped off by his parents at the start of his shift. Mike works a "h" hour shift. Write an expression for how much he makes in one night?

[Writes "h*5 = how much he makes"]

TUTOR That's right number.

Physics transcript:

T: [student name], I'd like you to read the problem carefully, and then tell me your strategy for solving this.

S: ok

[Pause 17 sec]

hmm.

[Pause 6 sec]

T: thinking out loud as much as possible is good

The Algebra transcript was made from a videotape of an algebra session where the tutor and the student are sitting in the same room. The Physics transcript was done from audio tapes of a phone conversation, where the tutor was seated in a room and students were seated in a different one. The tutor and students communicated with a conference phone that let only one person talk at a time. The main problem here is that within the same corpora, different formats were used to distinguish speakers, and linguistic annotations such as pauses in the algebra corpus were denoted by colon whereas in the physics corpus it was directly written inside brackets.

CSPA Corpus of Spoken Professional American-English (Athel 2002) includes transcripts conversations of various types occurring between 1994 and 1998. *For example:*

LANGER: Hello, I'm delighted to be here.

I have carefully read and heard about the University of Albany, the State University of New York. And I'm also the director of the National Research Center on English Learning and Achievement.

STRICKLAND: Her mother wrote the stances.

(Laughter)

KAPINUS: Dorothy, I might add also that Judith probably has more history with NAEP than just about that I know of, you know, NAEP and reading.

STRICKLAND: Yes, yes. We will really turn to you as a very important resource, Judith.

And we have a new member, Gloria Lopez Gutierrez.

And, Gloria, tell us a little bit about yourself.

In this transcript, we noticed long turns/monologues, and the transcripts were not "anonymised": speakers' last names were used.

The *TRAINS* Dialog Corpus (CISD 2002), a corpus of task-oriented spoken dialogue that has been used in several studies of human-human dialogue. *For example:*

utt9 : s: okay <sli> um

utt10 : what you'll have to do is you'll have to uh pick out an <sli> uh an engine
<sli> and schedule a train to do that

utt11 : u: okay <sli> um <sli> engine <sli> two

utt12 : s: + okay +

utt13 : u: + from + Elmira

utt14 : s: + mm-hm +

The main problem is dealing with extra-linguistic annotation like + and <sli>.

ICE-Singapore: International Corpus of English, Singapore English (Nelson 2002). *For example:*

<\$B>

<ICE-SIN: S1A-099#33:1:B>

How how are things otherwise

<ICE-SIN:S1A-099#34:1:B>

Are you okay

<\$A>

<ICE-SIN:S1A-099#35:1:A>

Uhm okay lah

<ICE-SIN:S1A-099#36:1:A>

Bearing up lah

<\$B>

<ICE-SIN:S1A-099#37:1:B>

Ah hah

<\$A>

<ICE-SIN:S1A-099#38:1:A>

Ya I mean I don't really feel comfortable talking about it over the phone so when I see you I'll tell you about it lah

This is a spoken dialog recorded through the phone. Problems here are: unconstrained conversations, a lot of linguistic annotation, and great variation in turn length is noticed.

<ICE-SIN:S1A-099#35:1:A> refers to text unit 35, in subtext 1, uttered by speaker A. ICE-SIN refers to corpus name ICE Singapore and S1A refers to academic spoken.

Mishler Book: Medical Interviews (Mishler 1985) is an example of a scanned text image, including a dialogue between a patient and his physician. The problem is that scanned image was not converted to text format and also uses extra-linguistic annotations.

We can summarise the DDC problems as follows:

No standard formats to distinguish between speakers.

Extra-linguistic annotations were used.

No standard format in using linguistic annotations.

Long turns and monologues.

Irregular turn taking (overlapping).

More than one speaker.

Scanned text-image not converted to text format.

For example, if the program reads the dialog transcript from the astronomy transcript in MICAS corpus, every pair of speakers will generate a new AIML category where the first speaker represents the pattern, and the second speaker represents the template after applying the filtering process. The AIML category generated is:

<category>

<pattern>CIRCUMPOLAR STARS. SO IF I KEEP MY POINTER THERE, EVERYTHING ELSE MOVES AND WE ALL GET SICK. AND WE GO BACKWARDS IN TIME. AND THAT'S EVEN MORE FUN.</pattern>

<template>make it go really really fast.</template>

</category>

The most significant problem with the DDC is the unstructured annotations used within its files. So each transcript has to be filtered and processes differently, which contradicts the generalization objective.

When re-engineering ALICE to a new domain or conversation style, the patterns and templates learnt from a training corpus are only a raw prototype: the chatbot and AIML files must be tested and revised in user trials. One of the main design considerations is how to plan the dialogue. A good dialogue design would mean less time testing and re-implementing AIML files.

We applied the same program version to a French dialogue corpus (Kerr 1983), which also required changing the pre processing text since it has its own specific annotations.

4. Software Implementation: Version 2

The second version of the program has a more general approach to finding the best match against user input from the learned dialogue. At first we decided to treat the problem of having more than two speakers within the dialogue corpora by 'recycling' each turn to be a pattern on one category and a template on the consecutive one. We used the same modules generated in the first version in order to read and pre-process the text. A restructuring module was added to evolve the program. The restructuring module searched the pattern 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. We then used an Afrikaans corpus (Van Rooy, 2002) to generate two new

versions of ALICE (Abu Shawar and Atwell, forthcoming): Afrikaans speaks only Afrikaans, and AVRA is bilingual and speaks both English and Afrikaans (most Afrikaans speakers are in fact bilingual). The bilingual version combined the standard ALICE AIML files that are written in English and the Afrikaans AIML file that is written just in Afrikaans. We used the <http://www.pandorabots.com/pandora> web-hosting service to make our chatbots available for use over the World Wide Web. User feedback from Afrikaans speakers suggested that we needed to extend the pattern-matching to enhance the responses generated.

To do this, we used the first word approach, based on the generalisation that the first word of an utterance may be a good clue to an appropriate response: if we cannot match the whole input utterance, then at least we can try matching just the first word. For each atomic pattern, we generated a default version that holds the first word followed by wildcard to match any text, and then associated it with the same atomic template. Unfortunately this approach still failed to satisfy our trial users, so we decided to use the most significant approach to augment the first word approach.

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 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. We extracted a local least frequent list from the Afrikaans corpus, and then compared it with each token in the pattern to specify the most significant word within that pattern. Four categories holding the most significant word were added to handle the positions of this word first, middle, last or alone. The feedback showed improvement in user satisfaction.

5. Conclusions

The implemented java program is able to convert the readable text extracted from a dialogue corpus to the ALICE format language. However a lot of problems raised using the corpus-based approach are illustrated in the different mark-up and annotation practices. To avoid these problems, the ideal training corpus must have the following characteristics: two speakers, structured format, short, obvious turns without overlapping, and without any unnecessary notes, expressions or other symbols that are not used when writing a text.

Even such "idealised" transcripts may still lead to a chatbot which does not seem entirely "natural": although we aim to mimic the natural conversation between humans, the chatbot is constrained to chatting via typing, and the way we write is different from the way we speak.

International research in NLP is dominated by work on English. NLP techniques and systems can be ported to other natural languages, but this is generally a labour-intensive task, requiring scarce computational and linguistic expertise; hence minority languages are poorly represented in NLP technology. We present an automated approach to porting an NLP technology, the AIML-based chatbot, to new languages, by using a corpus in the target language to retrain the chatbot. We have successfully automated production of chatbots talking different varieties of English, as well as French, and Afrikaans; and are developing further demonstrators in Spanish and Arabic.

6. References

ALICE 2002 *A.L.I.C.E AI Foundation* , <http://www.alicebot.org/> or <http://alicebot.franz.com/>

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. 2003 Using dialogue corpora to retrain a chatbot system, Proceedings of the Corpus Linguistics 2003 conference, pp681-690, Lancaster University UK.

AbuShawar B and Atwell E. 2003. Using the Corpus of Spoken Afrikaans to generate an Afrikaans chatbot. To appear in SALALS Journal of Southern African Linguistics and Applied Language Studies

Athel 2002 Corpus of Spoken Professional American-English: description,
<http://www.athel.com/corpdes.html>

CISD 2002 TRAINS Dialogue Corpus,
<http://www.cs.rochester.edu/research/cisd/resources/trains.html>

Kerr, B. (1983). *Minnesota Corpus*. University of Minnesota Graduate School, Minneapolis, USA.

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

MICASE 2002 *MICASE online homepage*, <http://www.hti.umich.edu/m/micase/>

Mishler E 1985 *The discourse of medicine: dialectics of medical interviews*, New Jersey, Ablex
<http://www-rcf.usc.edu/~billmann/diversity/Tr.5.1a.gif>

Nelson G 2002 *International Corpus of English: the Singapore Corpus user manual*, http://www-rcf.usc.edu/~billmann/diversity/ICE-SIN_Manual.PDF

Ringenberg M 2002 CIRCLE's tutorial archive <http://www.pitt.edu/~circle/Archive.htm>

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