A Corpus-based approach to generalising a chatbot system

Bayan Abu Shawar
University of Leeds
Leeds LS2 9JT, England
bshawar@comp.leeds.ac.uk

Eric Atwell
University of Leeds
Leeds LS2 9JT, England
eric@comp.leeds.ac.uk

Abstract: 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 French, and Afrikaans; and are developing further demonstrators in Spanish and Arabic.

Keywords: chatbot, dialogue, corpus, machine learning, English, French, Afrikáans, Arabic

Human machine conversation is a new technology to facilitate communication between users and computers via natural language. A chatbot is a conversational agent that interacts with users turn by turn using natural language. ALICE (http://www.alicebot.org/ , 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. 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), see http://www-rcf.usc.edu/~billmann , (Abu Shawar and Atwell 2003) to investigate the problems of utilising dialogue corpora. The DDC is a collection of links to different dialogue corpuses in different fields. These annotated texts are transcribed from recorded dialogues between more than two speakers. 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 fours 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.

The most significant problem with the DDC is the unstructured annotations used within its files. We applied the same program to a French dialogue corpus (Kerr 1983), which also required changing the pre processing text since it has its own specific annotations.

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 that 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 versions of ALICE: Afrikaana 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 Afrikaana 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.

To avoid the problems raised using corpus-based approach, 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.

Building French and Afrikaans versions of ALICE demonstrated the general approach. We propose to demonstrate the program further by developing other versions, including Spanish and Arabic chatbots.

References

Abu Shawar B, Atwell E. 2002. A comparison between ALICE and Elizabeth Chatbot systems, Technical report, School of Computing, University of Leeds.

Abu Shawar B, Atwell E. 2003. Using dialogue corpora to retrain a chatbot system, Proceedings of Corpus Linguistics 2003, pp681-690, Lancaster University.

Kerr, B. 1983. Minnesota Corpus. University of Minnesota Graduate School, Minneapolis.

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