

A Framework for Simulating and Evaluating Artificial Chatter Bot Conversations

Chayan Chakrabarti and George F. Luger

Department of Computer Science
University of New Mexico
Albuquerque, NM 87106 USA
cc@cs.unm.edu and luger@cs.unm.edu

Abstract

Chatter bots are computer programs that can simulate a conversation through text chat. Current chatter bots perform well in artificial conversations consisting of pairs of utterance exchanges like a question-answer session where the context switches with every pair. But they perform poorly in longer conversations where the context is maintained across several utterance exchanges. Existing approaches to artificial conversation generation focus on linguistic and grammatical modeling to generate individual sentence-level utterances. We present a framework that enables longer and more meaningful conversations by combining concepts of content representation and conversation semantics. We also present a metric for evaluating the conversations based on Grice's maxims, that form the central idea in the theory of pragmatics.

Introduction

Chatting with computer programs has come a long way since the pioneering artificial intelligence demonstrations like ELIZA (Weizenbaum 1966) and Parry (Colby 1973). Early chatter bots were designed as testing platforms for ideas in NLP and Computational Linguistics. Later efforts in chatter bot evolution focussed on making chatter bots more human-like. The Loebner Prize contest rewarded chatter bots considered to be the most human-like. Many famous chatter bots like A.L.I.C.E., Albert One, George, and Rosette were born out this effort (Saygin and Ciceklib 2002). Chatter bots have now evolved to a sufficient level of maturity and have found application in interactive games, as website navigation tools, and for simulating personal help desk assistants. It is estimated that by 2015, at least 50% of customer service will be realized via chatter bots, and this will result in cost reductions of up to 20% along with increased customer loyalty (Gartner 2012).

Contemporary chatter bots are sophisticated enough to leverage cutting age concepts like semantic parsing (O'Shea, Bandar, and Crockett 2009a) and sentiment analysis (Whitehead and Cavedon 2010). But have chatter bots reached the zenith of their ontogenesis, or are there opportunities to extend their current capabilities? We identify a key limitation

of current chatter bots, and present a framework to address it.

Observe snippets of an actual conversation with the chatter bot Rosette (<http://labs.telltalegames.com/rosette/>), which was the winner of the 2011 Loebner Prize.

Rosette: Good evening. My name is Rosette. What is your name?

Human: My name is Sally.

Rosette: Nice to meet you, Sally.

Human: Can you recommend a good store for designer clothing?

Rosette: I mostly wear jeans and a t-shirt.

Human: But I need a dress for a formal reception I am attending.

Rosette: Do you tend to like most things?

Human: No.

Rosette: OK. So, what do you do for a living?

Human: I am an accountant.

Rosette: What do you predict for the economy this year - bad or really bad?

Transcript 1: Conversation between Rosette and a human.

While Rosette seems articulate enough, the conversation clearly did not go the way Sally, the human, probably expected. So what exactly went wrong? Consider another snippet of conversation with an actual real-world customer service chatter bot that helps customers solve account access issues.

Chatter Bot: Hello. How can I help you today?

Customer: I am unable to transfer a balance from my holding account to my trading account.

Chatter Bot: You need to follow the instructions to reset your transaction password.

Customer: I do not want to reset my transaction password. I just reset it a week back.

ChatterBot: Transaction passwords need to be reset every fifteen days for account security.

Customer: I do not have a problem with my transaction password. I have a problem with my margin. I cannot transfer my holding balance.

Chatter Bot: Holding balance cannot be negative. Transactions that make the holding balance negative are automatically cancelled.

Transcript 2: Conversation between chatter bot and a customer trying to an electronic trading account issue.

Clearly, these state of the art chatter bots did not do well in producing satisfactory conversations. How do we quantify what is going wrong? If we observe the conversation closely, we notice a definite pattern. When the bots' response is evaluated only in relation to the immediate previous utterance by the human, they grade satisfactorily. It is only when evaluated on a longer, sustained conversation, that they grade poorly. They perform adequately in an isolated question-answer exchange. They even do well over a series of several consecutive question-answer pairs.

However, a series of question-answer pairs, or a series of one-to-one utterances, does not constitute a conversation. During a series of exchanges, the context switches from one pair to the next. But in most conversations, the context remains the same throughout the exchange of utterances. Contemporary chatter bots are unable to adhere to context in conversations. Our work aims to improve the conversational power of chatter bots. Instead of just being able to engage in question-answer exchanges, we want to design bots that are able to hold a longer conversation and more closely emulate a human-like conversation.

There are 3 ingredients for good chatter bot conversations, knowing what to say (content), knowing how to express it through a conversation (semantics), and having a standard benchmark to grade conversations (evaluation). There has been considerable progress in each of these areas through research in knowledge representation techniques (Beveridge

and Fox 2006), and conversation semantics (Oh and Rudnicky 2002; Lemon 2011). We present a well-defined standardized framework for simulating conversations and evaluating them. Unlike some of the endeavors mentioned earlier, we are not trying to solve the Turing test (Turing 1950). We are interested in designing chatter bots that can hold effective conversations in restricted well-defined domains.

Evaluation of Conversations

How do we differentiate between a satisfactory and unsatisfactory conversation? It was evident when reading Transcript 1 and Transcript 2, but how does one formally state this? It is important to have a standard, consistent metric to measure the quality of conversations as a means for comparison between different approaches, and a benchmark for seeking improvement. We borrow ideas from the theory of pragmatics to define some metrics to evaluate conversations.

Theory of Pragmatics

Pragmatics is a subfield of linguistics which studies the ways in which context contributes to meaning. Pragmatics encompasses speech act theory, conversational implicature, talk in interaction and other approaches to language behavior in philosophy, sociology, and linguistics (Mey 2001). It studies how the transmission of meaning depends not only on the linguistic knowledge (e.g., grammar, lexicon, etc.) of the speaker and listener, but also on the context of the utterance, knowledge about the status of those involved, and the inferred intent of the speaker. In this respect, pragmatics explains how language users are able to overcome apparent ambiguity, since meaning relies on the manner, place, time, etc. of an utterance (Grice 1957).

Pragmatics is a systematic way of explaining language use in context. It seeks to explain aspects of meaning which cannot be found in the plain sense of words or structures, as explained by semantics. As a field of language study, pragmatics' origins lie in philosophy of language and the American philosophical school of pragmatism. As a discipline within language science, its roots lie in the work of Paul Grice on conversational implicature and his cooperative principles.

Grice's Maxims

The cooperative principle describes how people interact with one another. As phrased by Paul Grice, who introduced it, "Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." (Grice 1989) Though phrased as a prescriptive command, the principle is intended as a description of how people normally behave in conversation. The cooperative principle can be divided into four maxims, called the Gricean maxims, describing specific rational principles observed by people who obey the cooperative principle; these principles enable effective communication. Grice proposed four conversational maxims that arise from the pragmatics of natural language. The Gricean Maxims are a way to explain the link between utterances and what is understood from them (Grice 1989).

Grice proposes that in ordinary conversation, speakers and hearers share a cooperative principle. Speakers shape their utterances to be understood by hearers. Grice analyzes cooperation as involving four maxims:

- * quality: speaker tells the truth that can be proved by adequate evidence
- * quantity: speaker is only as informative as required and not more or less
- * relation: response is relevant to topic of discussion
- * manner: speaker avoids ambiguity or obscurity, is direct and straightforward

It has been demonstrated that evaluating chatter bots using Grice’s cooperative maxims is an effective way to compare chatter bots competing for the Loebner prize (Saygin and Ciceklib 2002). The maxims provide a scoring matrix, against which each chatter bot can be graded for a specific criterion.

Simulation of Conversations

We next present a framework for modeling the content representation and semantic control aspects of a conversation. The key aspects for holding a conversation are knowing what to say that is both relevant to the context and within the domain (Chakrabarti and Luger 2012). The knowledge engine keeps the conversation in the right domain, while the conversation engine keeps it relevant to the context. These two modules handle distinct tasks in the conversation process. The chat interface module directly interfaces with the user.

Chat Interface

The high-level function of the Chat Interface (Figure 1) is to receive chat text from the user, pre-process this text and pass it on to the Knowledge Engine and the Conversation Engine. It then receives input back from the engines, and then transmits chat text back to the user. It has several sub-modules that facilitate this task. The Utterance Bucket is an interface that receives the chat text from the user and places the text into a buffer. The Stemmer module reduces the text to the root stems. This module examines the keywords, and detects the Speech Act associated with the keywords. The module has access to a set of keywords stored in a hash set. This module detects sentiment associated with the utterance. The standard set of bag of words pertaining to sentiments is used (Pang and Lee 2008). The topic module detects which topic keywords are present by referring to a hash set of topic keywords.

Knowledge Engine

The Knowledge Engine (Figure 2) supplies the content of the conversation. The two main content components of the conversation are information about the subject matter being discussed in this conversation organized using a Topic Hash Table and the particular speech act being followed in the conversation, which is organized using a Speech Act Hash Table. Each specific topic entry in the hash table consists of a set of context maps that encode specific information about the topic. The context maps are realized using several goal fulfillment maps.

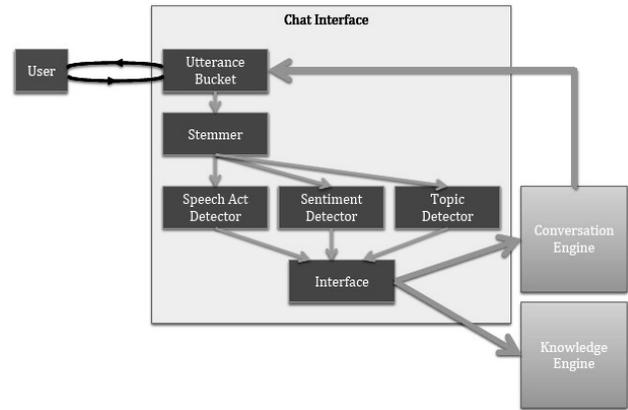


Figure 1: The Chat Interface module directly interfaces with the user.

A goal-fulfillment map is an effective way to represent the condensed knowledge in scripts. It is based on the conversation agent semantic framework proposed by O’Shea (O’Shea, Bandar, and Crockett 2009b; 2010). Engaging in dialogue with a user, the chatter bot is able to capture specific pieces of information to progress along a pre-defined network of contexts. In the example in Figure 3, where a chatter bot advises a customer of an electronic trading website about account issues, the contexts along the goal-fulfillment map expresses specific queries, which require specific answers in order for progression to be made along the designated route. Dialogue will traverse the goal-fulfillment map in a progression starting with the base context named *Initialize*. It is possible to revert to a previously visited context in the case of a misinterpreted line of input. The user can alert the chatter bot that there has been a misunderstanding. For example in following context, *Non Payment* aims to elicit the reason for non-payment of the margin fees; *Can Cover* identifies that the customer does have enough margin and thus goal-fulfillment is achieved; *Cannot Cover* aims to elicit why the customer doesn’t have sufficient margin; *Customer Status* identifies the status of the customer, and keeps following the map until goal-fulfillment is achieved.

Speech act theory asserts that with each utterance in a conversation, an action is performed by the speaker. These actions (or speech acts) are organized into conversations according to predefined patterns (GoldKuhl 2003). Winograd and Flores (Winograd and Flores 1986) show that conversation for action is an example of a pattern of speech acts organized together to create a specific type of conversation. We use a small subset of the 42 speech acts in the modified SWBD-DAMSL tag set (Jurafsky, Shriberg, and Biasca 1997). These are directly fed in to the Conversation Engine.

Conversation Engine

The Conversation Engine (Figure 4) supplies the semantic control for the chatter bot conversation. It contains two modules, the Probabilistic Finite State Automaton, which are

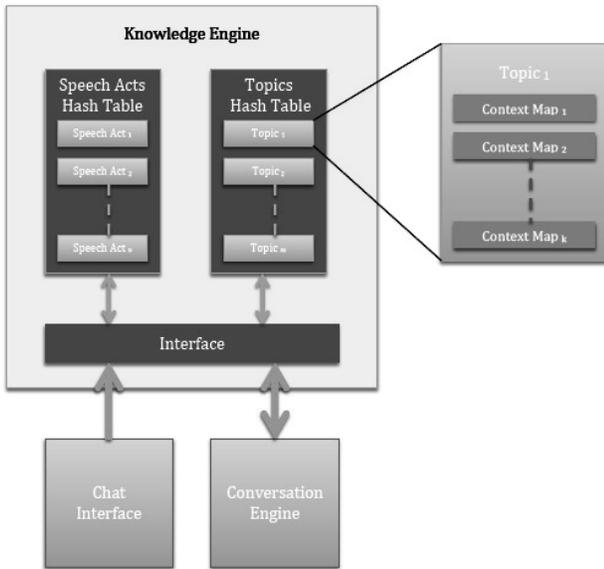


Figure 2: The Knowledge Engine contains the Speech Acts Hash Table, and the Topics Hash Table. Each Topic consists of Context Maps.

modeled from the Speech Act Hash Table in the Knowledge Engine, and the Conversation Planner, which is a list of strategies that engineer the conversation itself.

We have implemented a probabilistic finite state automaton (Figure 5), where states represent semantic states of the conversation, transitions represent classes of customer utterances, and an accepting state is the satisfaction state. We have identified five distinct semantic states in customer service conversations from a corpus of similar actual human conversations: *greeting* (includes small talk), *elicitation* (representative trying to understand the customer issue), *troubleshooting* (representative talks customer through several tests to gain more information), *resolution* (representative resolves the issue), and *satisfaction* (representative achieves customer satisfaction). The transition probabilities are learned from a corpus of human to human conversations in the same domain. The conversation control algorithm models the flow of the conversation with the transition probability matrix. The particular words or sentences in the utterance are obtained from the Topics Hash Table in the Knowledge Engine. The output of this module is sent to the Chat Interface Utterance Bucket through the interface.

Results

We show some sample conversations that our chatter bot is able to perform using our framework. We restricted our domain to customer service situations, in which a customer of an electronic trading website chats with a customer service chatter bot and seeks helps in resolving service related issues. These are actual traces of conversations that our chatter bot can perform when fed with simulated customer inputs. These conversations are graded by a panel

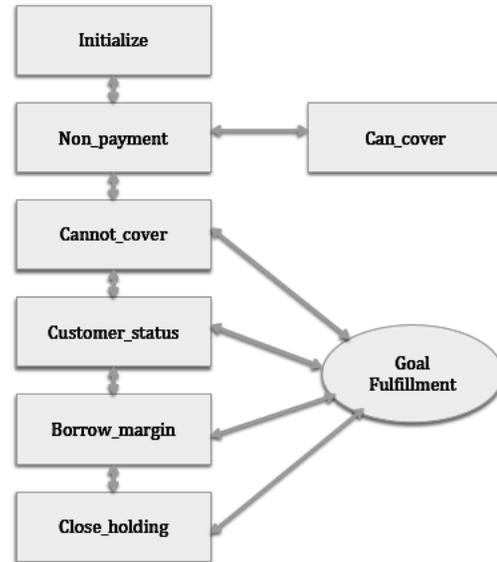


Figure 3: A Goal Fulfillment Map encodes the sequence of questions to be asked, whose responses will enable fulfill a specific goal during the conversation.

of human judges according to criterion specified by Grice's maxims. Since evaluating conversations is a subjective task, we have several experts evaluating a conversation according to the same predefined criteria and then collate the results together (Likert 1932).

Chatter Bot: Hello. How can I help you today?

Customer: My account shows an incorrect daily margin.

Chatter Bot: Is the margin higher or lower?

Customer: Lower.

ChatterBot: Do you have pending buy orders?

Customer: Yes, but I cancelled them before they were executed.

Chatter Bot: It takes 48 hours for cancel instructions to process.

Transcript 3: Conversation between our chatter bot and a customer trying to resolve a margin issue.

We observe that our chatter bot is able to successfully resolve the issue. Table 1 shows the evaluation of the conversation according to our evaluation framework.

Let us consider another conversation with our chatter bot

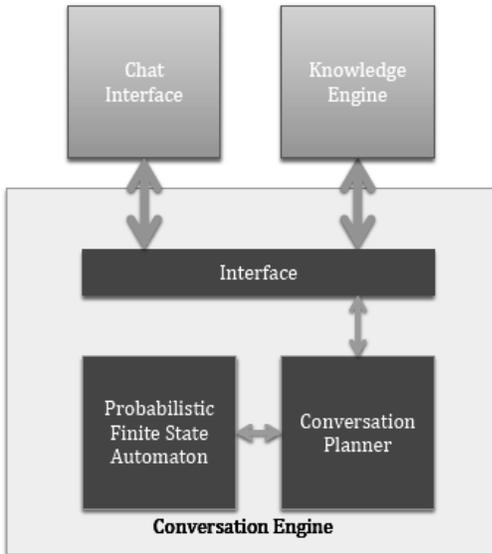


Figure 4: The Conversation Engine contains the Probabilistic Finite State Automaton, and the Conversation Planner.

Maxim	Result	Comments
Quality	Pass	Bot gives correct information for resolving the issue
Quantity	Pass	Bot's information is sufficient to resolve the issue
Relation	Pass	Bot correctly identifies issue, asks right questions, and gives advice relevant to the issue
Manner	Pass	Bot's answer is clear and unambiguous

Table 1: Evaluating Transcript 3

with a different issue.

Chatter Bot: Hello. How can I help you today?

Customer: I would like to open a new account for day trading. What are my options?

Chatter Bot: Do you have an existing demat account or would you like to open a new one?

Customer: I already have a demat account with my bank.

ChatterBot: What is the maximum amount of daily margin that you will require?

Customer: Not more than \$25,000.

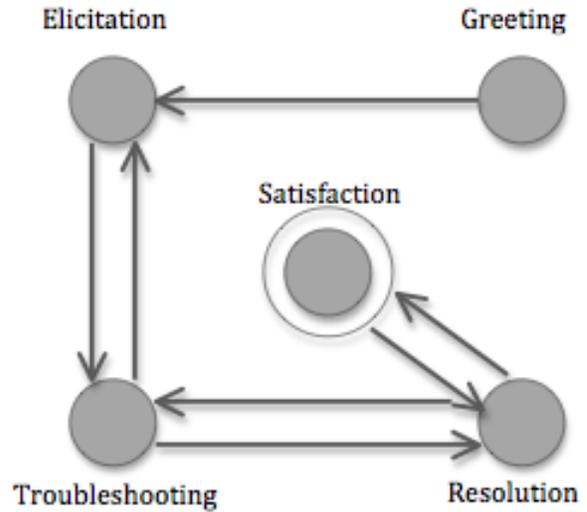


Figure 5: Finite State Automaton based representation of a conversation: the accepting states is *Satisfaction*. The other states represent various states of the conversation

Chatter Bot: Are you going to trade in one or more of futures and options, commodities, foreign exchange, and equities.

Customer: I am primarily interested in futures and options on commodities.

ChatterBot: The Intermediate Plan 2 account is best suited to your requirements. Daily margin of \$30,000. Commission of 1% per trade and immediate execution.

Transcript 4: Conversation between our chatter bot and a customer seeking information about a new account.

Once again we observe in Table 2 that our bot grades satisfactorily according to our evaluation framework.

Maxim	Result	Comments
Quality	Pass	Bot finds most suitable option
Quantity	Pass	Bot's seeks all required answer to make best decision
Relation	Pass	Bot correctly identifies query, seeks relevant information, and finds most suited option
Manner	Pass	Bot's queries are clear and unambiguous

Table 2: Evaluating Transcript 4

Each of these conversations grade well against Grice's

four maxims. They satisfy the *quality* maxim since the responses are factually correct. They satisfy the *quantity* maxim, since the information given is adequate and not superfluous. They satisfy the *relation* maxim since the responses are relevant to the context of the conversations. Finally they satisfy the *manner* maxim since the responses are unambiguous and do not introduce any doubt. Each of these conversations are also natural, similar to how a human customer service representative would communicate. We have simulated conversations on 50 issues in the domain of customer service for an electronic trading website, and managed to achieve good grades according to our evaluation criteria. We are in the process of deploying our framework in additional domains, and we are also experimenting with better conversation strategies and content representation techniques.

Conclusion

We have developed a robust framework for simulating and evaluating artificial chatter bot conversations. Our framework combines content representation techniques to provide the background knowledge for the chatter bot, and semantic control techniques using conversation strategies to enable the chatter bot to engage in more natural conversations. Our approach goes beyond lower level linguistic and grammatical modeling for utterance generation at the single sentence level granularity, and focusses on higher level conversation engineering.

Our evaluation framework is based on Grice's maxims and provides a standard benchmark for comparing and evaluating conversations that is consistent with the theory of pragmatics. Although the evaluation of individual conversations are subjective in nature, our framework provides a principled method for scoring them. We have shown that our chatter bots can perform well in targeted conversations.

The modular nature of our framework makes it suitable for plugging and experimenting with different approaches for the content representation and semantic control aspects of the conversation. Our evaluation framework also makes it possible to compare different approaches in a standardized manner. Our work attempts to aid in artificial conversation research by providing this standard framework for simulation and evaluation.

References

Beveridge, M., and Fox, J. 2006. Automatic generation of spoken dialogue from medical plans and ontologies. *Journal of Biomedical Informatics* 39:482–499.

Chakrabarti, C., and Luger, G. 2012. A semantic architecture for artificial conversations. In *The 13th International Symposium on Advanced Intelligent Systems*.

Colby, K. 1973. Idiolectic language-analysis for understanding doctor-patient dialogues. *IJCAI* 278–284.

Gartner. 2012. Organizations that integrate communities into customer support can realize cost reductions of up to 50 percent. *Gartner Customer 360 Summit 2012*.

GoldKuhl, G. 2003. Conversational analysis as a theoretical foundation for language action approaches? In Weigand, H.; GoldKuhl, G.; and de Moor, A., eds., *Proceedings of the 8th international working conference on the languageaction perspective on communication modelling*.

Grice, P. 1957. Meaning. *The Philosophical Review* 66(3).

Grice, P. 1989. *Studies in the Way of Words*. Harvard University Press.

Jurafsky, D.; Shriberg, L.; and Biasca, D. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. Technical report, University of Colorado - Boulder.

Lemon, O. 2011. Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. *Computer Speech and Language* 25:210–221.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 22(140):1–55.

Mey, J. L. 2001. *Pragmatics: An Introduction*. Oxford: Blackwell, 2 edition.

Oh, A. H., and Rudnicky, A. I. 2002. Stochastic natural language generation for spoken dialog systems. *Computer Speech and Language* 16:387–407.

O'Shea, K.; Bandar, Z.; and Crockett, K. 2009a. A semantic-based conversational agent framework. In *The 4th International Conference for Internet Technology and Secured Transactions (ICITST-2009)*, Technical Co- Sponsored by IEEE UK-RI Communications Chapter, 92–99.

O'Shea, K.; Bandar, Z.; and Crockett, K. 2009b. Towards a new generation of conversational agents using sentence similarity. *Advances in Electrical Engineering and Computational Science, Lecture Notes in Electrical Engineering* 39:505–514.

O'Shea, K.; Bandar, Z.; and Crockett, K. 2010. A conversational agent framework using semantic analysis. *International Journal of Intelligent Computing Research (IJICR)* 1(1/2).

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Saygin, A. P., and Ciceklib, I. 2002. Pragmatics in human-computer conversation. *Journal of Pragmatics* 34:227–258.

Turing, A. 1950. *Computing Machinery and Intelligence*. Oxford University Press.

Weizenbaum, J. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.

Whitehead, S., and Cavedon, L. 2010. Generating shifting sentiment for a conversational agent. In *NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 89–97. Association for Computational Linguistics.

Winograd, T., and Flores, F. 1986. *Understanding computers and cognition: A new foundation for design*. Norwood, New Jersey: Ablex Publishing Corporation.