Expert Systems with Applications 42 (2015) 6878–6897

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics

Chayan Chakrabarti *, George F. Luger

Computer Science Department MSC01 1130, 1 University of New Mexico, Albuquerque, NM 87131, USA

ARTICLE INFO

Article history:
Available online 2 May 2015

Keywords:
Artificial conversations
Computational conversations
Conversation engineering
Conversation semantics
Applied computational linguistics
Applied natural language processing

ABSTRACT

Chatter bots are software programs that engage in artificial conversations through a text-based input medium. They are extensively deployed in customer service applications. Existing approaches to artificial conversation generation emphasize grammatical and linguistic modeling techniques. They focus on generation of discrete sentence-level utterances. These approaches perform poorly in conversational situations requiring contextual continuity over a series of utterances. We present an approach that combines pragmatics with content semantics to generate artificial conversations in the customer service domain. A conversation is a process that adheres to well-defined semantic conventions and is contextually grounded in domain-specific knowledge. We model this using stochastic finite state machines, where the parameters of the model are learned from a corpus of human conversations. We present a specific set of criteria which we then use to evaluate the quality of artificial conversations in the customer service domain. We also compare chatter bot generated artificial conversations with human generated natural conversations in this domain.

1. Motivation: the need for artificial conversations

Many businesses have capitalized on the growth of text-based chat as an effective medium of communication and implemented their customer support operations through chat interfaces, where customers communicate with service representatives to resolve their issues. Analysts predict that by 2015, at least 50% of customer service will be performed by chatter bots, resulting in cost reductions of up to 20% and increased loyalty (Gartner, 2012). We illustrate a limitation of chatter bots using a snippet of actual conversation with a chatter bot that helps customers solve account access issues (Chakrabarti, 2014).

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

Customer: `I am unable to login to my account. I forgot my password.`

Chatter Bot: `Please click on the ''Forgot Password" link. You will be emailed a link that will help you change your password.`

Customer: `I already tried that. That link takes me to a form to reset password. But it requires me to enter my Customer Relationship Number. I do not have this number.`

Chatter Bot: `The Customer Relationship Number can be found in the ''Account Information" tab when you log into your account.`

Customer: `But I am unable to login to my account. I do not remember my password.`

Chatter Bot: `Please click on the ''Forgot Password" link. You will be emailed a link that will help you change your password.`

Clearly the chatter bot did not do well in this conversation. But is this part of a pattern? We use another example to further illustrate this subtle failure. In this example, a chatter bot deployed in customer service operations of an online electronic trading website tries to help a customer resolve an unallocated transaction, i.e., the customer's margin account does not have enough balance to service a pending buy order.

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

Customer: `I have a pending buy order and I am out of margin.`

Chatter Bot: `Buy orders are processed within 2 business days.`

* Corresponding author.
E-mail addresses: cc@cs.unm.edu (C. Chakrabarti), luger@cs.unm.edu (G.F. Luger).

http://dx.doi.org/10.1016/j.eswa.2015.04.067
0957-4174/

Customer: `But I am out of margin. I am concerned that my buy order will not execute.`

ChatterBot: `Margin can be restored by a new allocation request from your holding account. It takes 3 business days to execute.`

Customer: `But my buy order is already scheduled. I do not want to lose it.`

Chatter Bot: `Buy orders are processed within 2 business days.`

We observe that while the chatter bot can carry on a question–answer session satisfactorily, it is unable to tell if its response adequately addressed the customer's query or if the customer was confused by the response. Further examination of snippets of actual conversations in the customer service domain emphasize the chatter bot's lack of conversational capabilities (Chakrabarti & Luger, 2013; Chakrabarti, 2014).

We observe that state of the art chatter bots are more suited for question–answer type exchanges, where there is a sequence of several question–answer pairs. They are unable to hold a longer conversation, understand the conversation, gauge whether the conversation is going in the desired direction, and act on it. We expect the chatter bot to recognize the deadlock situation for the customer and resolve it, or transfer the conversation to a human representative if the solution is beyond its programmatic capabilities.

The next generation of chatter bots should be able spot such opportunities in a conversation, and act on them, either by disseminating relevant information, or by transferring the conversation to a human representative seamlessly. The missing element in current chatter bots is context awareness. In a series of question–answer exchanges, or pairwise utterances, 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 such a context in conversations.

2. Conversation and dialog engineering

Existing literature in conversation engineering consists of stochastic, syntactic, and semantic approaches. Stochastic approaches use statistical frameworks like Bayesian theory, Hidden Markov models, and n-gram modeling to construct individual sentences. Syntactic approaches use techniques from natural language processing and computational linguistics to parse and generate grammatical constructs of conversations. Semantic approaches use models of meaning from a knowledge structure to drive sentence analysis and construction.

One of the earliest conversational architectures was the GUS (Genial Understander System) (Bobrow et al., 1977), a virtual agent helping a customer make reservations. The system worked well on handling airline reservations, but it was not very intelligent. It could handle a restricted set of questions, and the domain knowledge of the question–answer sequence had to be encoded exactly in the same order in which the questions were asked.

The GALAXY Communicator system at MIT (Seneff et al., 1998; Polifroni & Seneff, 2000) is a client–server architecture for communicating online information including weather and flight information, and consists of database access, a speech synthesizer, a speech recognizer, and a language understanding engine. It has achieved good results in the travel reservation domain (Polifroni & Seneff, 2000). It can handle a wide range of conversations ranging from yes–no questions to answering complex queries (Filisko & Seneff, 2003). But it's not set up to build a knowledge base using facts, only in terms of anticipated questions (Filisko & Seneff, 2003).

The DARPA Communicator project (Levin et al., 2000) was an initiative to support advanced conversational capabilities including negotiation, plan optimization, and complex explanations. Evaluation metrics were the number of error messages, the mean system turn duration, the mean user turn duration, the number of system words to task end, the number of user words to task end, the mean response latency, and the total duration of the task (Walker, Hirschman, & Aberdeen, 2000).

The *Bayesian Receptionist* at Microsoft Inc (Horvitz & Paek, 2000), employed a set of Bayesian user models to interpret the goals of speakers given evidence gleaned from a natural language parsing of their utterances. Paek and Horovitz (2000) then demonstrated how conversations could be modeled as an inference and decision making problem under uncertainty.

State tracking is an important task in management of dialog systems. Several belief based state tracking architectures handle this problem using stochastic methods including generative and discriminative models (Deng et al., 2013). Some specialized techniques leverage dialogue structure in specific contexts to improve accuracy by encoding speech recognition patterns (Metallinou, Bohus, & Williams, 2013). Neural networks are used for deep-learning solutions (Henderson, Thomson, & Young, 2013). Partially Observable Markov Decisions Processes (POMDPs) are used to model conversations by improving ambiguity handling in changing domains (Gasic et al., 2013). Reinforcement learning is another successful approach (Rieser & Lemon, 2013).

Modern chatter bot implementations can effectively leverage computational linguistics techniques including semantic parsing (O'Shea, Bandar, & Crockett, 2009b) and sentiment analysis (Whitehead & Cavedon, 2010). Contemporary chatter bots perform very effectively in question–answer settings and similar utterance-exchange pair settings, where the contexts of the conversations are independent from one exchange to the next (Mauldin, 1994; Saygin & Ciceklib, 2002; Chakrabarti & Luger, 2012). However, they perform poorly in conversational situations where a specific context is maintained through a series of several utterance-exchange pairs. Existing customer service chatter bots are able to handle FAQ-type queries, but are unable to handle contexts that require a short conversation (Chakrabarti & Luger, 2012; Chakrabarti, 2014). Most chatter bot implementations focus on either content modeling or conversation semantics, or sub-aspects of these, or incorporate both of them together without making an explicit distinction. They do not distinguish the content required for the conversation from the semantics inherent in the conversation process. This leads to blind spots in the application, that we demonstrated in Section 1.

In this paper, we demonstrate a framework to overcome this limitation. Our chatter bot can generate artificial conversations in the customer service domain of electronic trading. These artificial conversations are of a better quality than modern chatter bots and do not have the limitations identified in Section 1. Section 3 introduces key concepts from the theory of speech acts, pragmatics, and knowledge representation which are leveraged in our framework. In Section 4, we define a set of metrics to evaluate artificial conversations. We describe our architecture in Section 5. In Section 6, we describe further details about our corpus and demonstrate how an artificial conversation is generated by our model. We demonstrate our experimental design and discuss results in Section 7. Finally, Section 8 presents conclusions and future work.

3. Modeling content and semantics

What is a conversation? There is no generally accepted definition, but the literature enumerates several characteristics. Conversations are based on a series of reactions to previous