AI behind conversational bots

Nalin Chhibber
Student ID: 20715659, MMath CS(HCI)

Abstract

The primary aim of this paper is to describe popular techniques that researchers have been using to build voice user interfaces and cover a sizable body of literature related to the area. Additionally, this paper replicates the results by (Hochreiter and Schmidhuber 1997) for Unstructured Multi-Turn Dialogue Systems. Based on the findings, the paper also introduces a new conversational agent: nBot. nBot is trained on a limited dataset of manually constructed sentences and build witt combination of both retrieval-based and generative techniques. It uses Long Short Term Memory to remember the context of conversation.

Introduction

Most of the existing modes of human-computer interaction(touch/type/click) require users to adapt with the interface(screen/keyboard/mouse). Voice user interfaces on the other hand, comes natural to humans and hence can be used as one of the most promising medium to engage them in a productive interaction. There has been a lot of optimism in the thought that near future will witness a rapid growth in human-computer-interaction using voice. This has not only led to an increased demand of conversational agents but also shown an increase in various chatbot development frameworks. Brands are increasingly using chatbots to engage their customers. Within just a couple years, we have seen a different evolution in the design of conversational agents from chatbots in Facebook Messenger, to Siri in iphones, to Microsoft's Cortana, Google Home and Amazon Alexa.

We are gradually incorporating technology in our lives, and with each step we try to simplify the interaction technique. Taking the example of a simple phone, we started using them with rotary dials, then replaced those with numeric buttons and moved on to using touch screens. Voice user interaction is the next logical step and we have already started seeing a glimpse of it in Siri and Google Assistant. While we cannot predict when Watson or Siri will start working like Jarvis, there is a still a lot that we can do with the normal scripted conversational agents.

Rest of the paper describes taxonomy of models, major challenges and related work around the design of conversational bots. It then details the implementation of nBot, followed by discussions on future work.

Taxonomy of models

Retrieval based vs Generative vs Pattern based

Challenge of building a conversational bot can be addressed using a retrival-based, generative or pattern-based model. Retrieval-based models are those which have a repository of pre-defined responses to answer a user utterance. Alicebot and Cleverbot are examples of such types. On the other hand, generative models are those which can generate responses they have never seen before. Microsoft's Tay bot is an example of generative model. Generative models are usually based on machine translation techniques but instead of translating from one language to another, they translate from input uttrance to output response. Both approaches have some obvious pros and cons. Retrieval based techniques dont make gramatical mistakes and are easier to train due to repository of handcrafted rules. However they may be unable to handle unseen cases for which no appropriate response exists. For the same reason these models cant refer back to contextual entity information. Generative models overcome this limitation and can refer back to entities in the input. However these models are harder to train, likely to make gramatical mistakes, produce inconsitant or irrelevant responses and require huge amounts of training data.

Another technique is to use pattern-based heuristics for selecting a response. In these techniques, when the agent receives a message, it goes through all the patterns until it finds a pattern which matches user utterance. If the match is found, the chatbot uses the corresponding template to generate a response. The problem with pattern-based heuristics is that patterns should be programmed manually which is not easy, especially if the agent has to correctly distinguish hundreds of intents. Users can express the same intent in many ways or use similar words in different contexts. Machine learning can help us train an intent classificaltion algorithm to pick patterns in the data. Most of the existing chatbot frameworks(wit.ai, dialogflow, Microsoft LUIS) are

based on this technique.

Open-domain vs Close-domain

In open-domain, conversations dont have a well defined goal or intention and can jump to any topic from dating to space travel to politics. Conversations on social media sites like Reddit and Twitter are typically open domain. In closed domain, space of possible input utterance is limited to a specific topic(education, casual conversation, technical support). It is realtively easy to model a conversational agent in close domain. Figure 1 illustrates relative difficulity in designing chatbots with different techniques.

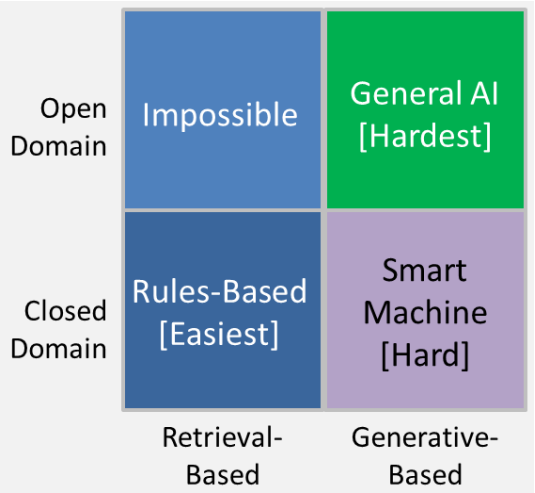

Figure 1: Chatbot Conversation Framework(Kojouharov 2016)

Online vs Batch Learning

Batch learning generates the best predictor by learning on the entire training data set at once. Online learning on the other hand uses new responses to update the best predictor for future responses at each step. Online learning can dynamically adapt to new patterns in the conversation as it is generated as a function of time. One problem with online learning is to control unwanted or potentially harmful information. One significant instance is when Microsoft's Twitter chatbot: Tay, started posting a deluge of incredibly racist messages in response to questions. Tay was designed to learn from conversations and get progressively "smarter" but online troublemakes and trolls persuaded it to blithely use racial slurs and even outright call for genocide (TechCrunch 2016). Therefore online learning should be assisted by other techniques to prevent them from going off the rails.

End-to-end vs Distributed Learning

Distributed learning systems require multiple stages of processing and follow a set pipeline for solving tasks at hand from feature extraction to learning the desired result. End-to-End learning on the other hand omits any hand-crafted intermediatery algorithms and directly learns the solution to a given problem from sampled dataset. This could involve concatenation of different neural networks(CNNs, RNNs) which are trained simultaneously. The idea is to let the network go from"raw-est" possible data to "final-most" output.

Challenges

Context Awareness

In real conversations, people don't start thinking from scratch every second to come come up with a new response. They build the conversation based on relevant concepts they might have discussed in past. Therefore, to produce sensible responses systems may need to incorporate both linguistic context and physical context. The most common approach is to embed the conversation into a vector, but doing that with long conversations is challenging in simple recurrent neural networks. Experiments by (Serban et al. 2016) and (Yao, Zweig, and Peng 2015) both go into that direction. Besides maintaining the context of topic, a smart conversational agent should also incorporate other kinds of contextual data such as date/time, location and information about a user.

Consistent Personality

When generating responses the agent should ideally produce consistent answers to semantically identical inputs. For example, the response to inputs like "How old are you?" and "What is your age?" should be same. This might sound simple, but incorporating such fixed knowledge or "personality" into models is very much a research problem. Since most of the models are trained on data from multiple different users, they tend to give inconsistent responses to the same user utterance. Works like (Li et al. 2016) has proven to be first steps into the direction of explicitly modeling a personality in a conversational bot.

Evaluation of Models

The ideal way to evaluate a conversational agent is to measure whether or not it is fulfilling its task.However such labels are expensive to obtain as there is no well-defined goal. Common metrics such as BLEU(Bilingual Evaluation Understudy) that are based on text matching and used for Machine Translation arent well suited because sensible responses can contain completely different words or phrases. In (Liu et al. 2016), authors have concluded that none of the commonly used metrics really correlate with human judgment.

Intention and Diversity

One of the prime issues with most fo the retrieval-based and generative models is lack of diversity in responses. Google's smart reply is a good example of this whose earlier version used to respond with "I love you" to almost anything. This is partly due to how these systems are trained, both in terms of data and in terms of actual training objective/algorithm. Some researchers have tried to artificially promote diversity through various objective functions (Li et al. 2015). Humans conversations are typically specific to the input and carry

an intention. Since generative models arent trained to have specific intentions they lack this kind of diversity.

Choice of corpus

A corpus is a principled collection of texts, written or spoken, which is stored on a computer. It must represent something and its merits are often judged based on how representative it is. There are many corpora available and some can be bought, some are free and some are not publicly available. There is no one corpus to suit all purposes and choice of corpora is really important while training a conversational agent for a specific purpose (O'keeffe, McCarthy, and Carter 2007). Some popular corpus that have been used to train conversational bots are as follows:

- NPS Chat Corpus: Consists of 10,567 posts in release 1.0. Part of python NLTK package.
- NUS Corpus: Collection of SMS messages compiled by NLP group at National University of Singapore.
- Ubuntu Dialogue Corpus: Consists of 1,000,000 examples based on chat logs from the Ubuntu channels on a public IRC network.
- Cornell Movie-Dialogs Corpus: Contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts: 220,579 conversational exchanges between 10,292 pairs of movie characters.
- Microsoft Research Social Media Conversation Corpus: Collection of 12,696 Tweet Ids representing 4,232 three-step conversational snippets extracted from Twitter logs.

For an effective research in dialog systems, corpus should have following qualities:

- Two-way (or dyadic) conversation, as opposed to multi-participant chat.
- Large number of conversations.
- Many conversations with several turns(more than 3).
- Task-specific domain, as opposed to chatbot systems.

Most of these requirements are satisfied by the Ubuntu Dialogue Corpus introducted by (Lowe et al. 2015). Following section describes the related work in the development of conversational agents but focus specifically on (Lowe et al. 2015), and report replicated results using their technique.

Related work

Initial work on goal driven dialogue systems primarily used rule-based systems with the distinction that machine learning techniques have been heavily used to classify the intention (or need) of the user, as well as to bridge the gap between text and speech(Serban et al. 2015). Research in this area started to take off during the mid 90s, when researchers began to formulate dialogue as a sequential decision making problem based on Markov decision processes. (Young et al. 2013) (Singh et al. 2000) (Pieraccini et al. 2009)

Implementation

Neural Networks

Recurrent Neural Networks

Long Short Term Memory

Discussion

A retrieval-based open domain system is obviously impossible because you can never handcraft enough responses to cover all cases. A generative open-domain system is almost Artificial General Intelligence (AGI) because it needs to handle all possible scenarios. Were very far away from that as well (but a lot of research is going on in that area). This leaves us with problems in restricted domains where both generative and retrieval based methods are appropriate. The longer the conversations and the more important the context, the more difficult the problem becomes.

Interaction modes Domain specific usage Corpus

References

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Kojouharov, S. 2016. Ultimate guide to leveraging nlp & machine learning for your chatbot. Accessed: 2017-12-9.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

O'keeffe, A.; McCarthy, M.; and Carter, R. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Pieraccini, R.; Suendermann, D.; Dayanidhi, K.; and Liscombe, J. 2009. Are we there yet? research in commercial spoken dialog systems. In *Text, Speech and Dialogue*, 3–13. Springer.

Serban, I. V.; Lowe, R.; Henderson, P.; Charlin, L.; and Pineau, J. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 3776–3784.

Singh, S. P.; Kearns, M. J.; Litman, D. J.; and Walker, M. A. 2000. Reinforcement learning for spoken dialogue systems. In *Advances in Neural Information Processing Systems*, 956–962.

TechCrunch. 2016. Microsoft silences its new ai bot tay, after twitter users teach it racism. Accessed: 2017-12-9.

Yao, K.; Zweig, G.; and Peng, B. 2015. Attention with intention for a neural network conversation model. *arXiv preprint arXiv:1510.08565*.

Young, S.; Gašić, M.; Thomson, B.; and Williams, J. D. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE* 101(5):1160–1179.