1 Introduction

The field of natural language processing, nlp, in computer science with the use of artificial intelligence, AI, and machine learning, ML, is perhaps one of the greatest accomplishments in the 20th and 21st century. From the first chatbot Eliza to the present-day Siri and Alexa, nlp has made its way into our lives in more ways than most of us realize. Some would argue that the use of AI in nlp rather than predicting the stock market would be a waste of time and that there are more profitable routes to pursue, however nlp is a promising area of computer science with much application to our daily routines. From helping handicap individuals communicate to translation to chatbots, nlp has promising innovations that we may be able to reach in the near future.

Chatbots will likely complement existing marketing, sales, and customer representative roles rather than provide firms with the ability to substitute human labor. Markets will likely experience shifts much like how automated teller machines, ATMs, impacted markets and teller duties at banks with their conception in the mid 20th century. Precisely, chatbots have the ability to complement such workers. While chatbots take away seemingly narrow conversations and provide users with responses that may take a low-level priority, this would allow the marketing, sales, and customer representative agents more time to develop a deeper level of connection with consumers to allow for a stronger connection between firms and consumers. The liberated time from these agents is likely to generate stronger brand loyalty and deeper roots with the firms using chatbots in their rendered goods and services.

Not all chatbots are the same, likewise the firms that implement them are likely to have different objectives when considering implementing them into their daily operations. Some are likely to incorporate a chatbot that is information seeking to assist customers sort through inventory (Deng, 2016). Another application may be more complex such as seeking the nearest point of interest, then defining its open hours, and then allowing a user to see availability or set reservations (Deng, 2016). The last type of bot, chatbots, seemingly tower over the other types of bots, allowing users to connect with the bot and engage in an inter-personal relationship (Deng, 2016). The different bots consist of varying degrees of complexity and exhibit shared challenges context, personality and intention comprehension.

2 Body

The varying degree of complexity in chatbots highly depends on their structure. They could be built around a recurrent neural net, a recursive neural net, or a convolutional neural net. Beyond the scope of the model itself, there are varying degrees of methodologies and libraries a developer can implement in order to achieve their goal. Packages like Textblob, Scikit Learn, Keras, and Tensorflow allow developers to work on any number of methods for their use cases.

Retrieval-based models (easier) use a repository of predefined responses and some kind of heuristic to pick an appropriate response based on the input and context. The heuristic could be as simple as a rule-based expression match, or as complex as an ensemble of Machine Learning classifiers. These systems dont generate any new text, they just pick a response from a fixed set (Britz, 2016). Generative models (harder) dont rely on pre-defined responses. They generate new responses from scratch. Generative models are typically based on Machine Translation techniques, but instead of translating from one language to another, we translate from an input to an output (response) (Britz, 2016). Both approaches have some obvious pros and cons. Due to the repository of handcrafted responses, retrieval-based methods dont make grammatical mistakes. However, they may be unable to handle unseen cases for which no appropriate predefined response exists. For the same reasons, these models cant refer back to contextual entity information like names mentioned earlier in the conversation. Generative models are smarter. They can refer back to entities in the input and give the impression that youre talking to a human. However, these models are hard to train, are quite likely to make grammatical mistakes (especially on longer sentences), and typically require huge amounts of training data (Britz, 2016). Deep Learning techniques can be used for both retrieval-based or generative models, but research seems to be moving into the generative direction. Deep Learning architectures like Sequence to Sequence are uniquely suited for generating text and researchers are hoping to make rapid progress in this area. However, were still at the early stages of building generative models that work reasonably well. Production systems are more likely to be retrieval-based for now (Britz, 2016)

Recurrent Neural Net

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer (Colah, 2015).

Sequences. Depending on your background you might be wondering: What makes Recurrent Networks so special? A glaring limitation of Vanilla Neural Networks (and also Convolutional Networks) is that their API is too constrained: they accept a fixed-sized vector as input (e.g. an image) and produce a fixed-sized vector as output (e.g. probabilities of different classes). Not only that: These models perform this mapping using a fixed amount of computational steps (e.g. the number of layers in the model). The core reason that recurrent nets are more exciting is that they allow us to operate over sequences of vectors: Sequences in the input, the output, or in the most general case both. A few examples may make this more concrete (Karpathy, 2015)

RNN computation. So how do these things work? At the core, RNNs have a deceptively simple API: They accept an input vector x and give you an output vector y. However, crucially this output vectors contents are influenced not only by the input you just fed in, but also on the entire history of inputs youve fed in in the past (Karpathy, 2015).

RNN character-level language models. That is, well give the RNN a huge chunk of text and ask it to model the probability distribution of the next character in the sequence given a sequence of previous characters (Karpathy, 2015).

In theory, RNNs are absolutely capable of handling such long-term dependencies. A human could carefully pick parameters for them to solve toy problems of this form. Sadly, in practice, RNNs dont seem to be able to learn them (Colah, 2015).

LTSM LONG SHORT TERM MEMORY

Long Short Term Memory networks usually just called LSTMs are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter Schmidhuber (1997), and were refined and popularized by many people in following work.1 They work tremendously well on a large variety of problems, and are now widely used (Colah, 2015).

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn (Colah, 2015).

Recursive Neural Net

Convolutional Neural Net

1. Using CNNs to analyze text is surprisingly effective. Our initial guess was that only RNNs (more specifically LSTMs) would be effective, given their ability to learn sequences and long-term dependencies, but weve learned that CNNs might have something to say about that. (G, 2016)

3 (

Common Challenges) Perhaps one of the greatest obstacles for chatbots is to be more human-like. Chatbots have been an area of interest and research since the mid 20th century and while innovation and research has led to many breakthroughs, they still have yet to obtain the human-like features that users need. The transition between narrow AI and general AI is a tremendous task for any application of AI, thus in an effort to make chatbots more practical there are certain areas of concentration that researchers need to emphasize on when developing such applications. Britz declares several areas which would vastly improve chatbots in their transition from narrow to general AI by incorporating context, coherent personality, and intention (2016). While narrow AI chatbots are more or less able to respond to generic and simple requests such as informing a user of the weather. This is likely to be determined by the grounds of the conversation dependent on whether or not the conversation is short or long and whether or not it is in an open or closed domain (Britz, 2016). Frankly, most short conversations are quite easy for chatbots and humans alike to handle. For example, how is the food here? and the response, superb is relatively straight forward, yet this is a short conversation in a closed domain. Change the context slightly and have someone iterate through an entire menu and ask about events near the diner and you have just added an extreme amount of complexity to the problem. A diner server might be able to inform one of the various menu items and events down the street, but the chatbot is likely to lose context of the situation through many questions and responses. One would think that it would not be so difficult to add context, yet some chatbots still lack the competency to fulfill such tasks. Adding another layer of depth to chatbots would allow them to incorporate context providing additional information to the user satisfying their utility of the chatbot. For any user, a single inconvenience is one too many, thus, for chatbots to excel in the future, adding context is imperative. The next challenge chatbots experience is the inability to be perceived as genuine or caring service agents. While research and development has improved these areas, it is still a challenge nonetheless. From the first chatbot Eliza to the more recent Amazon Alexa, development of personality has made remarkable strides such as being able to understand vulgarity in conversation and respond with politeness to deter such language as the system could identify it as offensive (Onlim, 2017). Finally, the next challenge is to incorporate intention. This is perhaps the hardest component to incorporate into chatbots. While chatbots are able to pick up on vulgarity as mentioned before, how do the chatbots recognize whether the language is directed towards the system and its responses or alternatively the ill-fated fact that the weather may be horrendous today or that the user finds that they had just missed the last showing of a movie they desperately wanted to see. These are incredible tasks and researchers and developers must focus on implementing these key features if chatbots will ever make it to a general artificial intelligence.

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 (Britz, 2016).

A.I. bots are now possible due to the recent huge advances in machine learning and A.I. These advances enable us to provide more and more automation for things we care about. The rise of deep learning in the past several years, particularly deep reinforcement learning (RL) in the past 1.5 years, makes effective use of the increasing amount of data and computing resources, boosting our ability to build computational models for the world environment and for any application domains relevant to our lives (Deng, 2016).

References

????

Britz, Denny. 2016. "Deep Learning for Chatbots, Part 1 Introduction." URL `http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/`.

Cho, Kyunghyun, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." *CoRR* abs/1406.1078. URL `http://arxiv.org/abs/1406.1078`.

Deng, Li. 2016. "How deep reinforcement learning can help chatbots." URL `https://venturebeat.com/2016/08/01/how-deep-reinforcement-learning-can-help-chatbots/`.

G, Abhinav. 2016. "Deep Learning Sentiment One Character at a T-i-m-e Gab41." URL `https://gab41.lab41.org/deep-learning-sentiment-one-character-at-a-t-i-m-e-6cd96e4f780`

Golle, Philippe. 2008. "Machine Learning Attacks Against the Asirra CAPTCHA." In *Proceedings of the 15th ACM Conference on Computer and Communications Security*, CCS '08. New York, NY, USA: ACM, 535–542. URL `http://doi.acm.org/10.1145/1455770.1455838`.

Karpathy, Andrej. 2015. "The Unreasonable Effectiveness of Recurrent Neural Networks." URL `http://karpathy.github.io/2015/05/21/rnn-effectiveness/`.

Karpathy, Andrej, Justin Johnson, and Fei-Fei Li. 2015. "Visualizing and Understanding Recurrent Networks." *CoRR* abs/1506.02078. URL `http://arxiv.org/abs/1506.02078`.

Olah, Christopher. 2015. "Understanding LSTM Networks." URL `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`.

Onlim. 2017. “The History of Chatbots.” URL `https://medium.com/@onlim`$_{c}om/the - history - of - chatbots - 2530dd3cdac5$.

Simonyan, Karen and Andrew Zisserman. 2014. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” *CoRR* abs/1409.1556. URL `http://arxiv.org/abs/1409.1556`.

Yao, Kaisheng, Trevor Cohn, Katerina Vylomova, Kevin Duh, and Chris Dyer. 2015. “Depth-Gated LSTM.” *CoRR* abs/1508.03790. URL `http://arxiv.org/abs/1508.03790`.