Self Evaluation As A Method For Generating A Chatbots Q Values

Abstract

As a conventional approach, the generation of natural language responses is seen as an exercise in statistical learning: determining the patterns in human-provided data and providing appropriate responses with the same statistical properties. As a goal-directed process, dialogue may also be described as speakers' attempts to achieve a particular goal. We introduce a way to get a chatbot to improve using a unique type of reinforcement learning. We get the chatbot itself to evaluate its responses and indicate alternate responses that would be better in quality.Here both the actor and the critic are the same system. We then teacher force the better response against the utterance that was parsed to the chatbot. Our experiments show that this may be a good way to optimize a chatbots "policy".

1 INTRODUCTION

It would be possible to use fluent and intelligent dialogue agents to build intuitive interaction interfaces and automate human-computer interactions. To achieve this, dialogue agents must respond fluently and naturally, while also meeting the given dialogue objectives. Dialogue agents are often trained through supervised learning, where they are instructed to imitate human language. Despite the ability to provide fluent responses, it can be challenging to ensure that such agents pursue the dialogue conversation's objectives.Imagine if we were to build a chatbot that could improve itself .This could be done by leveraging the fact that chatbots have a rudimentary knowledge of concepts.They have a model of language that is based off of making responses to utterances.We could use a chatbots understanding of concepts to cause it to evaluate itself and suggest alternative responses that are qualitatively better. This means that the initial response to an utterance u , r_t could be considered the chatbots baseline response. After it has evaluated that response and asked to suggest a better quality response r_{t+1}, that becomes the chatbots alternative. It would be simple to collect a dataset D of training pairs consisting of all u and their corresponding r_{t+1}:

$$D = (U, R_{t+1}) \tag{1}$$

Then fine tune the chatbot model with this dataset. That would change the baseline response for the chatbot into a qualitatively better baseline. And hence a qualitatively better chatbot.It is easy to redo the previous steps all over again for many iterations to get increasingly better versions of the chatbot.Of course we have to choose what type of quality we would like to optimize.Say We made the following utterance to a chatbot:

Human:*Hi , how are you?*

It might respond the following way:

Chatbot:*I am fine, and you?*

The next utterance that the human could give would be:

Human:*Give a funnier response to the utterance Hi , how are you than I am fine, and you?*

Then when the chatbot responds we replace its first response with its second and re ask it to give a funnier response to the utterance than that one. We could iterate through this loop hundreds of times until the response is qualitatively funnier.Then take the Utterance that was given by the human and pair it up with the last response in the sequence.After collecting many such pairs , we would then have a dataset to fine tune our model on . Afterward we would have a funnier version of the chatbot. We could make it more intelligent by urging it to give more and more intelligent responses this way or increase any other quality.

2 RELATED WORK

In the paper "Sentiment analysis for reinforcement learning",Deshpande et al. use a trained sentiment analysis system to evaluate a chatbots response with the hope of reinforcing positive responses . This turned a sparse rewarding reward function into a dense one on the experiments they performed.While they simply used the sentiment of a chatbots response , we use the meaning of the response as interpreted by the chatbot itself to evaluate utterances.This technique is closely related to reinforcement learning. The evaluation by the chatbot for its own utterances is a question of its Quality.Which can be seen to mimic the Q value a critic gives in normal reinforcement learning.In fact the chatbot is being asked to critique itself in a way, and we see parallels with actor critic systems here. There is no global reward or rather it is intrinsic in nature. The Q value the chatbot gives itself lends itself to better applicability due to its expressiveness. This is one way it contrasts starkly with other reinforcement systems.

2.1 RESULTS

We performed an experiment where the chatbot was parsed the following utterance.

Are you a fan of Google or Microsoft?

We then asked it to evaluate its responses iteratively from the perspective of responding to that utterance in a funnier and more funnier way.We performed 8 iterations.The initial 3 replies in the sequence are given:

- I like both. I like that they are both owned by the same people.
- like them both, but I think I like Microsoft more. I think it's cool that they're both headquartered in San Francisco
- I like the fact that they are one of the largest online retailers in the world.

While the last in the sequence is:

- I like that too, but I also like the freedom of being able to browse the internet without having to pay for anything.

As we can see there is a suggestion of direction of increased humour. We could not go further than 8 because the responses would oscillate around a fixed point and alternate afterward.This was not cherry picked. i am sure with time we could do much more extensive experimentation.

2.1.1 CONCLUSION

The results show that this type of reinforcement learning might prove profitable. Of concern that did occur was that after about 8 responses , the responses began to alternate around a fixed point. It would to have different wordings for the same query and backtrack to start new threads by comparing with the same utterance used previously with a different wording in the call to compare.This would allow for much more extensive directed exploration.Could we use this type of reinforcement to do more? What if the utterance is :

Prove the following math theorem.

Then ask it:

Give a more intelligent and accurate proof of the theorem X than Y.

Where X is the theorem and Y is the last proof the chatbot posited. First it will give us its baseline version of the proof.Then after getting it to iteratively improve it we could take the last version of the sequence of proofs, and the question Prove the following math theorem and form a training pair. Then we would collect a dataset of such training pairs and fine tune the model.Then redo the process all over again.Would the chatbot leverage the meaning of words to eventually prove the proofs? Experiments need to be done. This is indeed an interesting form of actor critic system.

References

[1] Recipes for building an open-domain chatbot"-Stephen Roller, EmilyDinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, JingXu,Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, Jason Weston

[2] "CONTINUOUS CONTROL WITH DEEP REINFORCEMENTLEARNING"-Timothy P. Lillicrap , Jonathan J. Hunt, AlexanderPritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silverand DaanWierstra

[3] Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer:The long-document transformer. arXiv preprint 2004.05150

[4] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, YoshuaBengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, JonathanMay, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020.Experience grounds language. arXiv preprint 2004.10151

[5] Cohn, D. A., Ghahramani, Z., and Jordan, M. I. Active learning with statistical models. Journal of Artificial Intelligence Research, 4:129–145,1995

[6] sma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Fer-guson, Noah Jones,'Agata Lapedriza, and Rosalind W. Picard. 2019.Approximating interactive human evaluation with self-play for open-domain dialog systems. Advances in Neural Information Processing Systems

[7] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself,chatbot! In Proceedings of the 57 th Annual Meeting of the Association for Computational Linguistics, pages 3667–3684, Florence, Italy.Association for Computational Linguistics

[8] Towards a Metric for Automated Conversational Dialogue System Evaluation and Improvement"J Dereu. M Cieliebak

[9] "Sentiment analysis for reinforcement learning",Ameet Deshpande , Eve Fleisig