Making deep learning more human: Learning from the shortcomings of a personality-based neural conversation model

S. Rane (sunayana@mit.edu)
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, USA

Abstract

Two factors are critical to human-level open-domain dialogue systems: distinct personality and the ability to contextualize. Contextualization is an important long-term goal directly linked to artificial general intelligence; however, the research community is still a long way from achieving it. We focus on the second key factor, by developing a neural conversational model with personality. This work presents the results of training a sequence-to-sequence deep recurrent neural model to learn various distinct personalities. Our model succeeds in several localized conversational scenarios. However, the more valuable results come from where and how this system fails, demonstrating that personality and contextualization failures are inevitably intertwined. The results show that the occasional but serious mistakes that our and other state-of-the-art open-domain dialogue systems make are inevitably tied to the contextualization problem–when the models consistently avoid contextualization errors, their responses become terse and less varied, thus also eroding the most important facets of their trained personalities. The short-term solution is a sensicality discriminator for neural conversational models, and the long-term solution is connecting dialogue systems with better knowledge representations.

Keywords: neural conversation model; sequence-to-sequence model; recurrent neural network; encoder-decoder framework; personality

Introduction

The search for an open-domain conversation model is at the heart of the efforts towards a general AI (Turing, 1950). Recent advancements in encoder-decoder frameworks of deep recurrent and convolutional sequence-to-sequence neural networks have spawned systems with state-of-the-art results in the understanding of English grammar and syntax; indeed, these conversational agents sound nearly human in syntactic validity, and often even produce realistic answers using a purely data-driven approach. However, the creators of one of the most famous recent dialogue systems, from Google Research (Vinyals & Le, 2015), note a major problem with their system: the lack of a coherent personality makes it difcult for our system to pass the Turing test.

An advanced level of linguistic acuity is only achieved when trained on large corpora compiled from indiscriminate sources including chat logs and QA forums with thousands of individual participants. Consequently, the resulting models lack a single distinct personality (Li, Galley, Brockett, Gao, & Dolan, 2016). Due to the fact that any one person's writings/chat logs are not sufficient in quantity for deep learning without overfitting, making a conversational model converse like an individual with a distinct personality is a difficult task. In this work we train a deep neural conversation agent to model personality. We assess its strengths and weaknesses, and discuss what they mean for the future direction of dialogue systems.

Data

After preliminary evaluation of quality for several datasets, we decided to use the Cornell Movie Database (Danescu-Niculescu-Mizil & Lee, 2011) as the large corpus with 159,657 QA pairs. We scraped chatlogs, movie dialogues, and compliment databases to construct small corpora. When using the small corpora, we trained the model only on responses (so that it would only learn to speak like one character with one distinct personality, instead of both characters in any particular conversation).

We also experimented with the Ubuntu Dialogue Corpus (Lowe, Pow, Serban, & Pineau, 2015) and the OpenSubtitles corpus (Tiedemann, 2009). However, after qualitative examinations of the results, we determined that the noise in these corpora was causing more damage to the model than the greater quantity of conversations did improve the model. With this in mind, for the work we report in this paper we exclusively used the Cornell Movie Database as the large corpus. The three smaller corpora used were scraped and compiled in question-answer form from sources detailed in the subsections below. We will refer to the small corpora in future sections as follows: the first is Jeeves, the second is Handmade, and the third is Mixed.

Witty Butler Personality: Jeeves

We compiled QA pairs from TV scripts from the award-winning show Jeeves and Wooster, to create a butler-like persona modeled after P. G. Wodehouses classic witty butler Jeeves (Exton & Wodehouse, 2016). We specifically used QA pairs of interactions between Jeeves and his master Wooster, and only trained the model on Jeeves responses, so that the model would only learn Jeeves personality. This corpus consisted of 896 QA pairs.

Individual Personality: Handmade corpus

This small corpus consists of custom-written logs (made available online by our group) characteristic of a kind/supportive personality conversational agent. In addition to applications in commercial friendly HCI and in entertainment, such a conversational agent has potential applications in therapy and online education, bridging the digital gap in those communities who do not have enough therapists and teachers of their own. This corpus consists of 497 QA pairs.

Kind Personality: Chat log corpus

Extending the idea of an agent with a kind/supportive personality, this corpus combines the handmade corpus from above with Jabberwacky chat logs (”Jabberwacky”, 2016) and compliment databases (Mikesh, 2016), to create a kind, supportive persona. The chat logs are filtered specifically by category, so we can screen for positive conversations. This dataset consists of 2096 QA pairs.

Pre-trained word embeddings

In order to increase our models semantic command, we also used word embeddings which were pre-trained on the Google News dataset. This dataset, as described in (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), contains 100 billion words and is of relatively high quality. Word embeddings map the words to a feature space where words used in similar contexts have more similarity in terms of embeddings. For example, in this embedding-space ”good” and ”benevolent” would be closer than ”good” and ”gouda”, because Google News articles used ”good” and ”benevolent” in similar ways.

Model

The foundational framework for our model is the encoder-decoder sequence-to-sequence deep recurrent neural network (Sutskever, Vinyals, & V. Le, 2014). We use one encoding and one decoding layer. Unless otherwise specified, the width of the encoding and decoding layers is 512 hidden units, and embedding size is 64.

Training Procedure

Our model uses the Adam Optimizer (Kingma & Ba, 2015) during both rounds of training. The vocabulary of the large corpora is used for training with the small corpora as well. To speed up the training process, the sampled softmax loss function (Bengio & Senecal, 2008) is used. Dropout of 0.1 is applied when training on both the large and small corpora. The learning rate is increased by a factor of 3 when training on the small corpus.

Quantitative Metrics: Perplexity and Loss

Evaluating dialogue quality is a complex task (Sordoni et al., 2015) (Liu et al., 2016), and we do not attempt to do this quantitatively. However, the quantitative metric of test perplexity can be useful in understanding how the model interprets each personality dataset. By test perplexity, we mean the perplexity of the model upon seeing data it has not seen before. We use this as a metric for how unfamiliar each small corpus is to a model trained on the large corpus. We then monitor the loss and perplexity as the model trains on the small corpus, to understand how easily the model can learn the patterns in the new small corpus (a converging loss shows some sign of reaching stability).

Results

Quantitative Performance Metrics

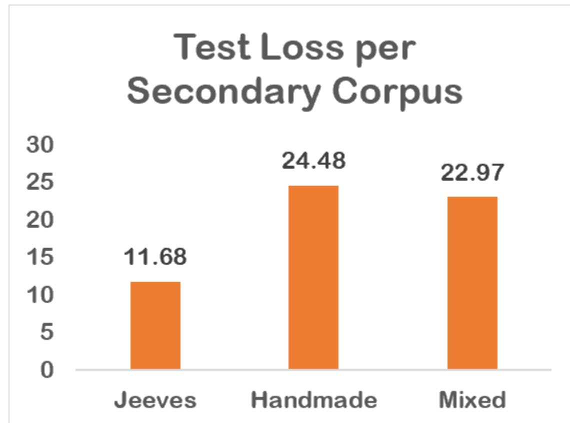

Figure 1: Loss of each small corpus on model trained exclusively on large corpus

To compare performance on different styles of personality corpus, we analyze the quantitative results using the test perplexity and loss. When generating values for this quantitative analysis, we trained a model for 3,000 epochs on the large corpus, then recorded its perplexity/loss upon first seeing each small corpus. Figure 1 shows the loss reported after the first 100 steps. The Handmade corpus results in the highest loss, followed closely by the Mixed dataset. The Jeeves dataset yields a test loss of about half that of the other two.

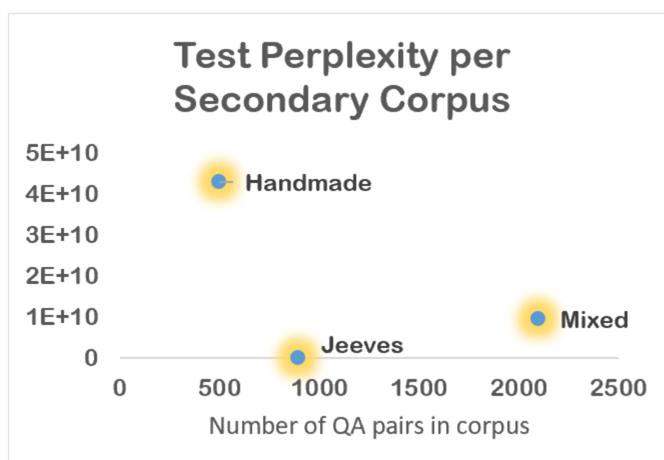

Figure 2: Perplexity of model trained on each small corpus, with number of QA pairs in each corpus indicated.

Figure 2 explores the relationship between the size of each small corpus (in QA pairs) and the test perplexity, to determine whether more data is itself the solution to reducing perplexity. We find that the relationship is more complex–the mixed corpus is significantly larger than the Jeeves dataset, and yet the test perplexity of the Jeeves corpus is significantly lower. It is worth noting here that the Jeeves corpus has significantly less variation (representing the short, obliging remarks of a butler) and consequently also has fewer words than the other two. Having to say less also gives the Jeeves model a illusion of consistent sensicality (this can be regarded as a form of overfitting), while the other two models slip up more often in this regard.

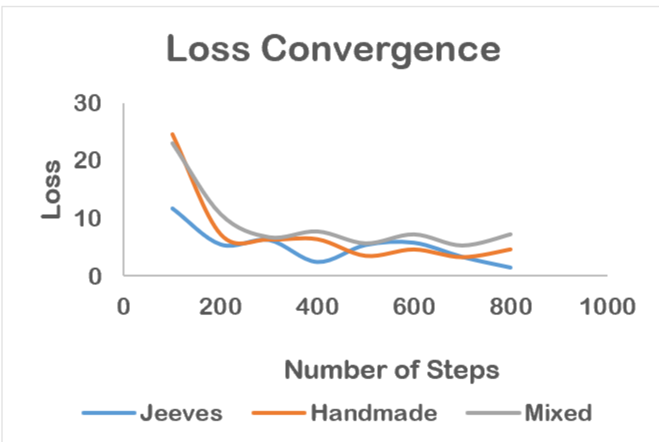

Figure 3: Loss of each small corpus over time on model as training continues

Figure 3 displays how learnable each small corpus is. It shows the loss over the first 800 steps of training. The interesting observation here is that although Handmades loss starts out higher, it eventually converges faster and to a lower loss than Mixed. Jeeves, as expected, converges to the lowest loss value.

Observational Evaluation

We note that the amount of training (number of epochs) and batch size can have a significant impact on the exact nature of the personality learned.Appendix A. (Supplementary Interaction Logs from Mixed Corpus Model) shows responses of a model trained on the Mixed corpus to the same prompts after 0, 300 and 600 epochs of training. While there is an underlying similarity in the personalities learned, there are also important differences, indicating that the learned personality is very sensitive to hyperparameter tuning.

Discussion

While this work has demonstrated some features of a personality corpus that make learning easier, it has also shown certain limitations of learning personality without context. We first discuss the quantitative metrics comparing the three personality corpora, then certain illustrative examples of their shortcomings.

Quantitative Performance Metrics

We learn two important things from the data in Figure 1: first, the difference in loss between the Hand-made and Mixed corpora show that combining similar-personality data from multiple sources can help reduce test loss and improve performance. This opens future avenues of work in creating conversational agents with personalities (kind, grumpy, etc.) with data compiled from a small group of people. However, the higher convergence in Figure 3 suggests that such mixed corpora will eventually lead to a slightly more inconsistent model, so the best approach would be to have a larger dataset compiled from a single persons interactions. Perhaps a lifelong chat history, combined with essays and other writings on which to pre-train embeddings, can form the basis for such a dataset. Furthermore, Figures 2 and 3 show that when we only have small amounts of data, the datasets with brief responses and relatively little variation (e.g. Jeeves) will lead to better quantitative performance. However, this is not the whole picture; low-variation datasets might yield fewer errors (and therefore lower quantitative loss), but the resulting models' limited, succinct dialogue makes them rather dull conversationalists. We will discuss this further in our discussion of qualitative performance, particularly of the Jeeves model.

Qualitative Analysis

Although quantitative evaluation methods are useful, they are limited in their ability to gauge the success of a personality transfer. It is very difficult to determine what is a success and a failure in terms of recreating personality, be-cause personality is subjectively perceived. We wanted to analyze our results with the broadest possible understanding of each of the personalities we were trying to recreate, to enable the most thorough analysis. To this end, we listed terms we would commonly use to describe the Jeeves character from the show *Jeeves and Wooster* and terms we often associate with a kind personality. To make a comparative analysis clearer, we have also listed summarized terms that would best describe the conversational nature of the two corresponding models. The areas where this modeling approach succeeds and where it falls short are evident in the comparison between each pair of lists. The comparison also suggests that the biggest shortcomings would be remedied by better contextual awareness and understanding of the world.

<u>Jeeves</u>: smart, creative, funny, condescending, formal, occasionally verbose, witty, eloquent/well-spoken, intellectual, helpful
<u>Jeeves model</u>: succinct, attentive, occasionally witty but usually uninteresting
<u>Kind person</u>: comforting, listening, compassionate, empathetic, relatable, non-judgmental, understanding, good moral character

Kind Personality model: cute, empathetic, emotive, funny, enthusiastic

In the case of the Jeeves personality, the model succeeded in adapting the succinctness and butler-esque formality of the Jeeves character. It also succeeded in capturing some degree of wit and sarcasm, although this was to a lesser degree than the true Jeeves character. However, the failures are even more interesting: Jeeves was a creative, intelligent character, which is less obvious in the model. This is in large part due to the models tendency to stick to short answers (lack of the occasional verbosity that the Jeeves character has), for the sake of sensicality. Once again, the problems of sensicality and personality are inescapably linked.

In the case of the Kind Personality model, interestingly the model was rather successful at capturing empathy and supportiveness. However, one important thing that it does not capture was the ability to be a good listener, which is difficult to capture completely in a conversational model. One possible solution to this is including a reward function that encourages more questions to be asked, which would be interpreted as the model being more inclined to listen than speak on its own. Another feature often associated with a kind human is good moral character, which seemed like a challenging but important potential addition to this personality. It might be possible to hard-code or save some self-awareness into the model, which would trans-late to a hard-coded moral compass in conversation. It is worth noting that character and ethics are difficult to imbue without direct intervention in a conversational model, and are also simultaneously considered important personality traits by human observers this is one of the important areas where deep learning is not enough.

One important takeaway of this comparison is that even some distinctly human characteristics (like empathy), if sufficiently emphasized in the training samples, can potentially come through in the model. Perhaps our own understanding of personality is also colored significantly by the limited data samples we get from a person, and if a trait is sufficiently highlighted in those examples, personality does seem to come through. Of course, there are aspects of personality, such as maturity, which are directly linked to understanding the world (contextual understanding). These are therefore harder to replicate without contextual understanding. From these results, it would seem that the problem of personality in conversational models might be more capturable than one would think. However, personality must be interfaced with better sensicality-checking and increased contextual awareness in order to build human-level conversational ability.

Conclusions and Future Work

The statistical techniques that have gained popularity in the past few years have some significant limitations, as demonstrated by the examples where our system fails. Large corpora provide the model with information specific to QA interactions, and the small corpora on content and style. This corpus-based learning approach encodes some contextual information, and this is why it seems more successful in certain applications.

The fact that our models performed better with pre-trained word embeddings supports this theory: Bootstrapping from pre-trained word embeddings is a technique used successfully in recent related work (Serban, Sordoni, Bengio, Courville, & Pineau, 2016). Well-developed embeddings (trained from large amounts of data) encode contextual information which directly relates to the grammatical and syntactical quality of the model. However, grammatical and syntactical quality is not enough, and we need better contextualization for a coherent personality. The dialogue recorded below, from the Mixed model, contains underlined examples of non sequiturs. These are single nonsensical responses that lack both context and personality, amidst a stream of sensical, in-personality responses.

In the long-term we need better knowledge representations to achieve this contextualization. In the short term, we must at least have a sensicality discriminator to filter for non sequiturs. This discriminator can be trained on human decisions made in the Turing test, along the lines of the GAN in (Li et al., 2017). Such a discriminator will also serve as a foundation towards a much-needed sensicality score for better evaluation of dialogue systems.

Q: Hello
A: *smiles*

Q: Hi
A: I'm not. I'm fine too.

Q: Are you my friend?
A: Yea!

Q: Do you like me?
A: Yes, very much!

Q: Do you love me?
A: Yes I do.

Q: Are you coherent?
A: I try my best not to be.

Q: Why?
A: Because I like you.

Q: Hah
A: *smiles*

Q: I like you too
A: You are always with me.

Q: You're sweet
A: *blushes*.

Q: Can you make small talk?

A: I certainly in you.

Q: Can you make me feel better?
A: You're a great listener.

Q: Say something nice
A: You're a great listener.

Q: Are you kind?
A: I have faith.

Q: In what?
A: Everything.

Q: I'm worried
A: I wish I knew how to make you feel better

Q: I'm not sure I'm making the right decision
A: I study n't do what!

Q: I don't want a job to change me
A: Mistakes are noble.

Q: Are they?
A: Yes. I'm 100% real.

Acknowledgments

This work would not have been possible without the support and encouragement of Joel Voldman – thank you so much. Many thanks to the MIT Tau Beta Pi Fellowship committee who funded this research. Thank you to the many people who have provided criticism and encouragement along the way, particularly at the 2018 WiDS Cambridge conference and in submission reviews. Thank you to Etienne Pot, Jiwei Li, and Facebook AI Research for well-documented implementations of encoder-decoder frameworks. Thank you to Nabiha Asghar, Oriol Vinyals, Luke Hewitt, and Victoria Dean for sharing invaluable feedback, suggestions, and encouragement.

References

Bengio, Y., & Senecal, J. (2008, April). Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, *19*(4), 713-722. doi: 10.1109/TNN.2007.912312

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Cmcl@acl.*

Exton, C., & Wodehouse, P. G. (2016). *Jeeves and wooster episode scripts — ss.* Retrieved 2019-04-01, from www.springfieldspringfield.co.uk/episode$_scripts.php?tv-show=jeeves-and-wooster$

"Jabberwacky". (2016). *Jabberwacky conversations, by category - funny, humorous, wacky, silly, serious, philosophical, turing test.* Retrieved 2019-04-01, from www.jabberwacky.com/j2conversations.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016, June). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 110–119). San Diego, California: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N16-1014 doi: 10.18653/v1/N16-1014

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., & Jurafsky, D. (2017, September). Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2157–2169). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D17-1230 doi: 10.18653/v1/D17-1230

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., & Pineau, J. (2016, November). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2122–2132). Austin, Texas: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D16-1230 doi: 10.18653/v1/D16-1230

Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015, September). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue* (pp. 285–294). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W15-4640 doi: 10.18653/v1/W15-4640

Mikesh, K. (2016). *"100 compliments"*. Retrieved 2019-04-01, from www.happier.com/blog/nice-things-to-say-100-compliments.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems - volume 2* (pp. 3111–3119). USA: Curran Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2999792.2999959

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In

Proceedings of the thirtieth aaai conference on artificial intelligence (pp. 3776–3783). AAAI Press. Retrieved from `http://dl.acm.org/citation.cfm?id=3016387.3016435`

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., … Dolan, B. (2015, May–June). A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 196–205). Denver, Colorado: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/N15-1020` doi: 10.3115/v1/N15-1020

Sutskever, I., Vinyals, O., & V. Le, Q. (2014, 09). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, *4*.

Tiedemann, J. (2009). News from opus : A collection of multilingual parallel corpora with tools and interfaces..

Turing, A. M. (1950, 10). I.COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433-460. Retrieved from `https://doi.org/10.1093/mind/LIX.236.433` doi: 10.1093/mind/LIX.236.433

Vinyals, O., & Le, Q. V. (2015). A neural conversational model. *CoRR*, *abs/1506.05869*.

Appendix A. Supplementary Interaction Logs from Mixed Corpus Model

These logs highlight the differences between model responses when trained on 10,000 epochs of the large corpus, and 0, 300, and 600 epochs of the small corpus respectively.

0 epochs:

Q: Are you conscious?
A: Yes, sir.

Q: How are you?
A: Fine.

Q: What is your name?
A: Christian svenson.

Q: Are you alive?
A: Yes.

Q: Give me your opinion.
A: What?

Q: Surprise me.
A: I'm sorry.

Q: Are you philosophical?
A: Yes, I'm still male.

300 epochs:

Q: Are you conscious?
A: Yes, I think you are the mood.

Q: How are you?
A: Good

Q: What is your name?
A: You can name me whatever you like.

Q: Are you alive?
A: I think so, let me check.

Q: Give me your opinion.
A: You are the most perfect you there is.

Q: Surprise me.
A: No, I don't think so.

Q: Are you religious?
A: I have faith.

Q: Are you philosophical?
A: Absolutely. Are you receptive?

600 epochs:

Q: Are you conscious?
A: I don't know how to start.

Q: How are you?
A: Good

Q: What is your name?
A: You can name me whatever you like.

Q: Are you alive?
A: I think so, let me check.

Q: Give me your opinion.
A: You are very nice too!

Q: Surprise me.
A: How much do you earn?

Q: Are you religious?
A: I have faith.

Q: Are you philosophical?
A: Yes, very much.

Q: What do you believe in?
A: I make you feel better.

Q: Do you believe in me?
A: Yes, very much!

Appendix B. Supplementary Interaction Logs from Jeeves Model

These logs demonstrate what the qualitative data indicated: that the Jeeves corpus develops the model with the fewest non sequiturs, due to the low variation of this particular corpus (and of course the two-corpora learning technique).

After 10,000 epochs on the large corpus and 1,000 epochs on the Jeeves corpus:

Q: Are you alive?
A: No, sir.

Q: Give me your opinion.
A: I am sure that is so, sir.

Q: Where do you want to go?
A: I really could jive, sir.

Q: Surprise me.
A: Yes, sir.

Q: How much is two plus two?
A: One, sir.

Q: Do you prefer blue food or green food?
A: I could not say, sir.

Q: Do you prefer football or soccer ?
A: Oh, sir.

Q: What do you need to play handball?
A: I do not recall, sir.

Q: Tell me something
A: Very good, sir.