Google AI Blog

The latest from Google Research

Towards a Conversational Agent that Can Chat About… Anything

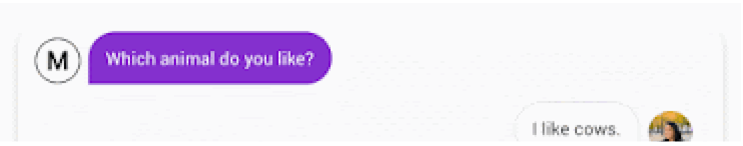
Tuesday, January 28, 2020

Posted by Daniel Adiwardana, Senior Research Engineer, and Thang Luong, Senior Research Scientist, Google Research, Brain Team

Modern conversational agents (chatbots) tend to be highly specialized — they perform well as long as users don't stray too far from their expected usage. To better handle a wide variety of conversational topics, open-domain dialog research explores a complementary approach attempting to develop a chatbot that is not specialized but can still chat about virtually anything a user wants. Besides being a fascinating research problem, such a conversational agent could lead to many interesting applications, such as further humanizing computer interactions, improving foreign language practice, and making relatable interactive movie and videogame characters.

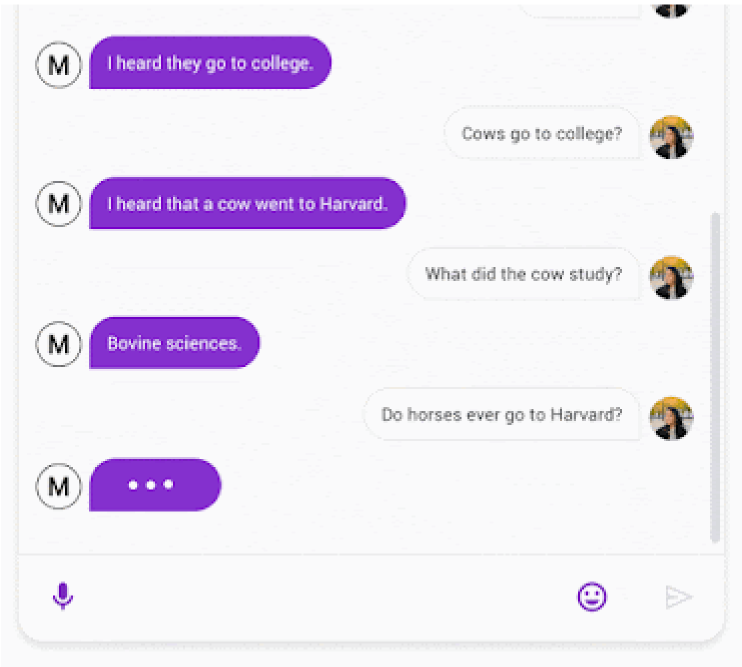
However, current open-domain chatbots have a critical flaw — they often don't make sense. They sometimes say things that are inconsistent with what has been said so far, or lack common sense and basic knowledge about the world. Moreover, chatbots often give responses that are not specific to the current context. For example, "I don't know," is a sensible response to any question, but it's not specific. Current chatbots do this much more often than people because it covers many possible user inputs.

In "Towards a Human-like Open-Domain Chatbot", we present Meena, a 2.6 billion parameter end-to-end trained neural conversational model. We show that Meena can conduct conversations that are more sensible and specific than existing state-of-the-art chatbots. Such improvements are reflected through a new human evaluation metric that we propose for open-domain chatbots, called Sensibleness and Specificity Average (SSA), which captures basic, but important attributes for human conversation. Remarkably, we demonstrate that perplexity, an automatic metric that is readily available to any neural conversational models, highly correlates with SSA.



Which animal do you like?

I like cows.



M: I heard they go to college.

Cows go to college?

M: I heard that a cow went to Harvard.

What did the cow study?

M: Bovine sciences.

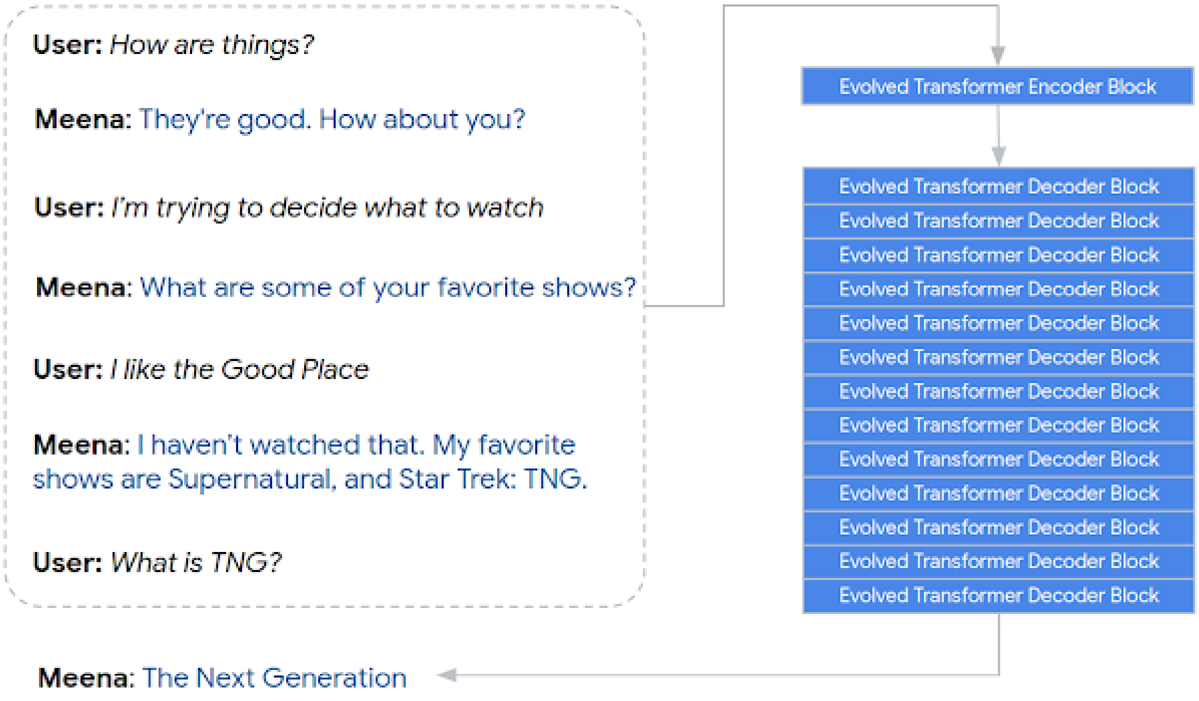
Do horses ever go to Harvard?

M: •••

A chat between Meena (**left**) and a person (**right**).

Meena
Meena is an end-to-end, neural conversational model that learns to respond sensibly to a given conversational context. The training objective is to minimize perplexity, the uncertainty of predicting the next token (in this case, the next word in a conversation). At its heart lies the Evolved Transformer seq2seq architecture, a Transformer architecture discovered by evolutionary neural architecture search to improve perplexity.

Meena has a single Evolved Transformer encoder block and 13 Evolved Transformer decoder blocks, as illustrated below. The encoder is responsible for processing the conversation context to help Meena understand what has already been said in the conversation. The decoder then uses that information to formulate an actual response. Through tuning the hyper-parameters, we discovered that a more powerful decoder was the key to higher conversational quality.



User: *How are things?*

Meena: They're good. How about you?

User: *I'm trying to decide what to watch*

Meena: What are some of your favorite shows?

User: *I like the Good Place*

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*

Meena: The Next Generation

Evolved Transformer Encoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block

Example of Meena encoding a 7-turn conversation context and generating a response, “The Next Generation”.

Conversations used for training are organized as tree threads, where each reply in the thread is viewed as one conversation turn. We extract each conversation training example, with seven turns of context, as one path through a tree thread. We choose seven as a good balance between having long enough context to train a conversational model and fitting models within memory constraints (longer contexts take more memory).

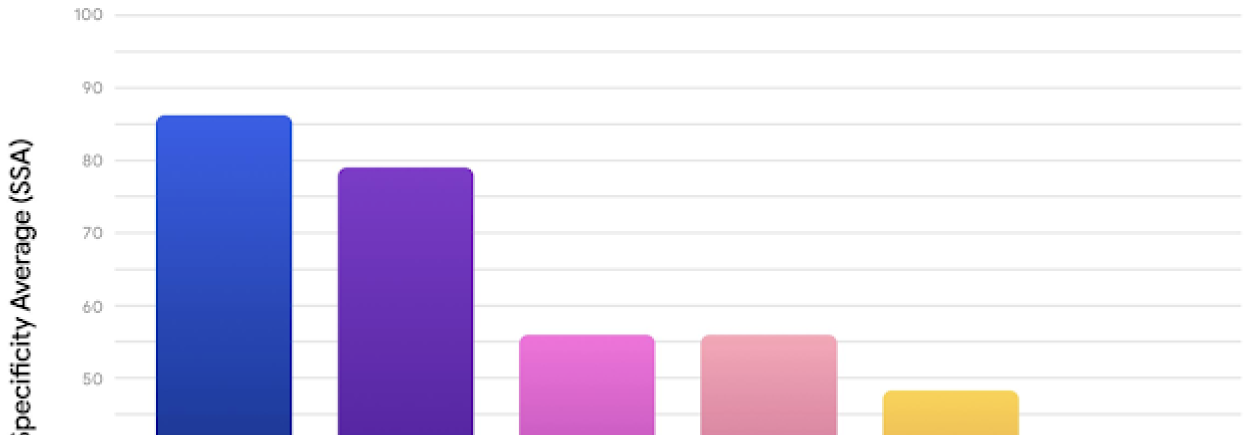
The Meena model has 2.6 billion parameters and is trained on 341 GB of text, filtered from public domain social media conversations. Compared to an existing state-of-the-art generative model, OpenAI GPT-2, Meena has 1.7x greater model capacity and was trained on 8.5x more data.

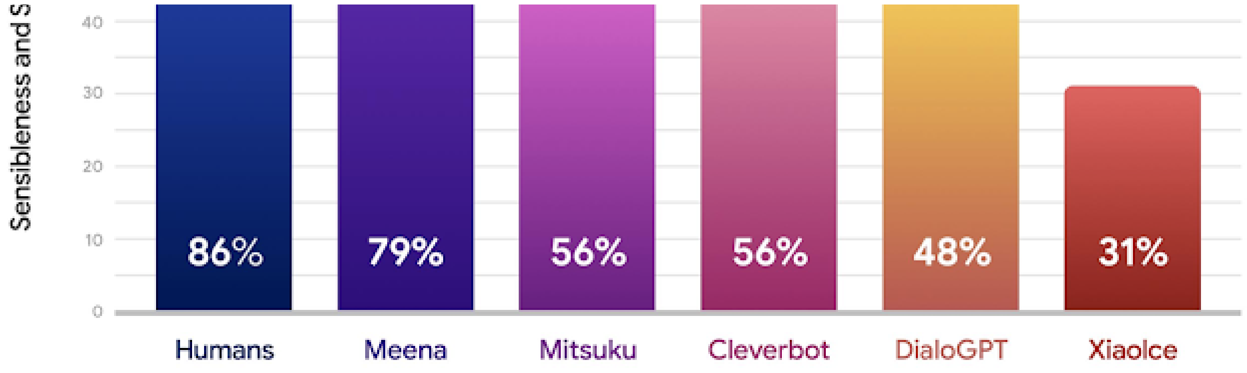
Human Evaluation Metric: Sensibleness and Specificity Average (SSA)

Existing human evaluation metrics for chatbot quality tend to be complex and do not yield consistent agreement between reviewers. This motivated us to design a new human evaluation metric, the Sensibleness and Specificity Average (SSA), which captures basic, but important attributes for natural conversations.

To compute SSA, we crowd-sourced free-form conversation with the chatbots being tested — Meena and other well-known open-domain chatbots, notably, Mitsuku, Cleverbot, XiaoIce, and DialoGPT. In order to ensure consistency between evaluations, each conversation starts with the same greeting, “*Hi!*”. For each utterance, the crowd workers answer two questions, “does it make sense?” and “is it specific?”. The evaluator is asked to use common sense to judge if a response is completely reasonable in context. If anything seems off — confusing, illogical, out of context, or factually wrong — then it should be rated as, “does not make sense”. If the response makes sense, the utterance is then assessed to determine if it is specific to the given context. For example, if *A* says, “*I love tennis*,” and *B* responds, “*That’s nice*,” then the utterance should be marked, “not specific”. That reply could be used in dozens of different contexts. But if *B* responds, “*Me too, I can’t get enough of Roger Federer!*” then it is marked as “specific”, since it relates closely to what is being discussed.

For each chatbot, we collect between 1600 and 2400 individual conversation turns through about 100 conversations. Each model response is labeled by crowdworkers to indicate if it is sensible and specific. The sensibleness of a chatbot is the fraction of responses labeled “sensible”, and specificity is the fraction of responses that are marked “specific”. The average of these two is the SSA score. The results below demonstrate that Meena does much better than existing state-of-the-art chatbots by large margins in terms of SSA scores, and is closing the gap with human performance.

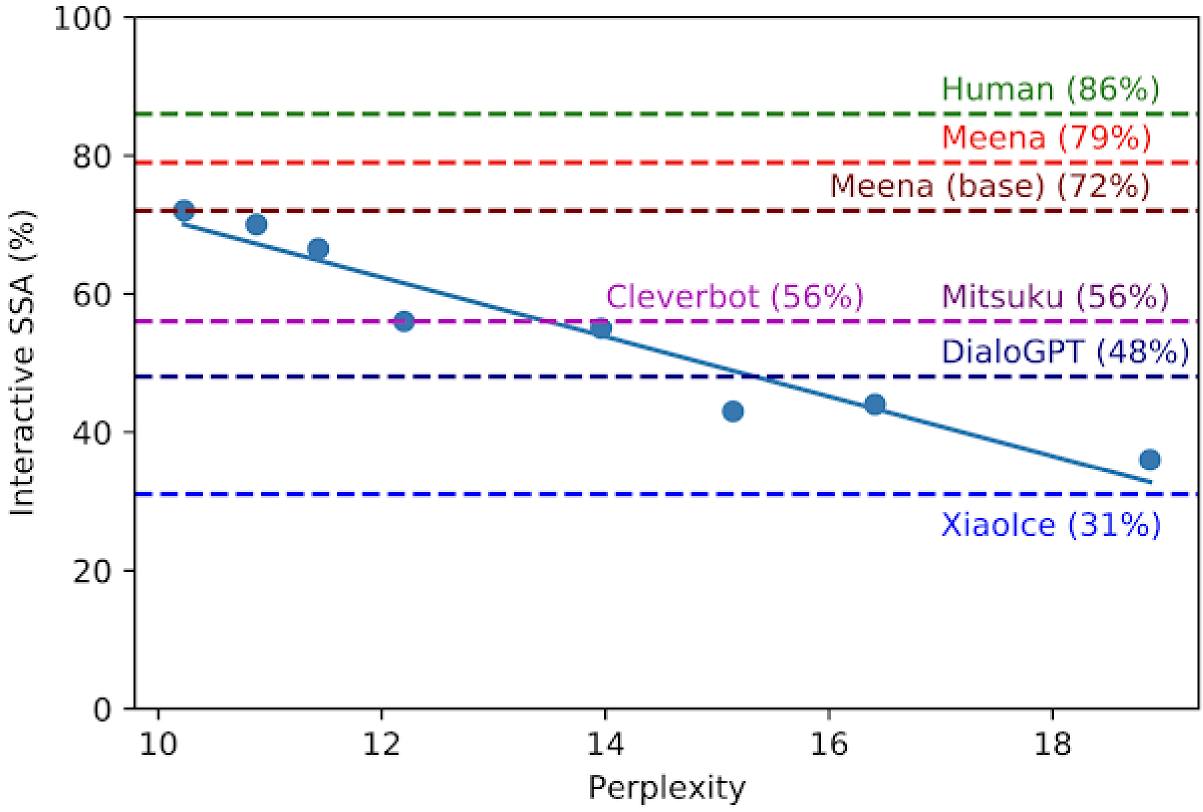




Meena Sensibleness and Specificity Average (SSA) compared with that of humans, Mitsuku, Cleverbot, XiaoIce, and DialoGPT.

Automatic Metric: Perplexity
Researchers have long sought for an automatic evaluation metric that correlates with more accurate, human evaluation. Doing so would enable faster development of dialogue models, but to date, finding such an automatic metric has been challenging. Surprisingly, in our work, we discover that perplexity, an automatic metric that is readily available to any neural seq2seq model, exhibits a strong correlation with human evaluation, such as the SSA value. Perplexity measures the uncertainty of a language model. The lower the perplexity, the more confident the model is in generating the next token (character, subword, or word). Conceptually, perplexity represents the number of choices the model is trying to choose from when producing the next token.

During development, we benchmarked eight different model versions with varying hyperparameters and architectures, such as the number of layers, attention heads, total training steps, whether we use Evolved Transformer or regular Transformer, and whether we train with hard labels or with distillation. As illustrated in the figure below, the lower the perplexity, the better the SSA score for the model, with a strong correlation coefficient (R^2 = 0.93).



Interactive SSA vs. Perplexity. Each blue dot is a different version of the Meena model. A regression line is plotted demonstrating the strong correlation between SSA and perplexity. Dotted lines correspond to SSA performance of humans, other bots, Meena (base), our end-to-end trained model, and finally full Meena with filtering mechanism and tuned decoding.

Our best end-to-end trained Meena model, referred to as Meena (base), achieves a perplexity of 10.2 (smaller is better) and that translates to an SSA score of 72%. Compared to the SSA scores achieved by other chabots, our SSA score of 72% is not far from the 86% SSA achieved by the average person. The full version of Meena, which has a filtering mechanism and tuned decoding, further advances the SSA score to 79%.

Future Research & Challenges

As advocated previously, we will continue our goal of lowering the perplexity of neural conversational models through improvements in algorithms, architectures, data, and compute.

While we have focused solely on sensibleness and specificity in this work, other attributes such as personality and factuality are also worth considering in subsequent works. Also, tackling safety and bias in the models is a key focus area for us, and given the challenges related to this, we are not currently releasing an external research demo. We are evaluating the risks and benefits associated with externalizing the model checkpoint, however, and may choose to make it available in the coming months to help advance research in this area.

Acknowledgements

Several members contributed immensely to this project: David So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu. Also, thanks to Quoc Le, Samy Bengio, and Christine Robson for their leadership support. Thanks to the people who gave feedback on drafts of the paper: Anna Goldie, Abigail See, YizheZhang, Lauren Kunze, Steve Worswick, Jianfeng Gao, Daphne Ippolito, Scott Roy, Ilya Sutskever, Tatsu Hashimoto, Dan Jurafsky, Dilek Hakkani-tur, Noam Shazeer, Gabriel Bender, Prajit Ramachandran, Rami Al-Rfou, Michael Fink, Mingxing Tan, Maarten Bosma, and Adams Yu. Also thanks to the many volunteers who helped collect conversations with each other and with various chatbots. Finally thanks to Noam Shazeer, Rami Al-Rfou, Khoa Vo, Trieu H. Trinh, Ni Yan, Kyu Jin Hwang and the Google Brain team for their help with the project.