LaMDA, Meena and the current state of Chatbots

NLP Report

January 12, 2021

Lachlan Todd
Christine Roesch
Jan-Matthias Jetter
Sami Hassane
Nahira El Yabroudi
Manuel Lehoux
Jose Antonio Garcia
Jaime Quiros

Introduction

Chatbots have seen explosive growth in recent years and are at the forefront of cutting edge research in the domain of Natural Language Processing (NLP) and Artificial Intelligence. The research conducted to build better language models has wide implications for the NLP field ranging from machine translation to generative language models and open-domain dialogue systems. The innovations in this field have already impacted a large cross-section of the industry with companies adopting chatbots for customer support services and internal operations.

Many companies have only been able to implement what is known as a closed-domain system, which focuses on a specific domain. For example, a travel site's chatbot might be able to help you book a flight, but it can't discuss the weather or have a conversation. In contrast, open-domain systems are built to understand conversation in any domain and as such have been much trickier to develop. Human conversations are complex and unpredictable. We can be serious and humorous, figurative and literal, or informational and creative in the span of a few minutes switching between a variety of topics. Capturing this essence of human-like conversation is the next frontier for conversational models. Human-like open-domain chatbots have the potential to disrupt entire corporate functions and enable entire new product categories. Recent research has achieved astounding results for open-domain chatbots.

In May 2021, Google released LaMDA, a breakthrough conversational bot, able to engage in free-flowing and open-ended dialogue that is highly sensitive to the context of the conversation. Through short demos available online, one can observe how LaMDA can engage in the chaotic nature of a conversation, by jumping from one topic to the next with great ease and without the need to finetune the models' parameter's for every new topic. This has been a big challenge for open-domain chatbots in the past. It is still early days to see which commercial applications will be built on top of LaMDA. However, it's easy to see how this revolutionary innovation in language modeling has the potential to change how humans consult the internet and interact with digital applications.

Unlike many other chatbots, LaMDA was trained on dialogue and has been fine-tuned to have the following qualities:

1. Specificity - provides specific/detailed responses given the conversational context
2. Factuality - provides information based on facts
3. Interestingness - provides responses that are insightful, unexpected or witty
4. Sensibleness - provides responses that show emotional qualities and personality

LaMDA is unique as it has been trained to especially excel in all of the above mentioned qualities: specificity, factuality, interestingness and sensibleness. While developing LaMDA, Google discovered that specificity and sensibility correlate very strongly with perplexity, an automatic metric used to evaluate the human-like quality of language models (perplexity measures how uncertain a model is about the predictions it makes the next word ought to be). LaMDA has been trained with the objective to specifically minimize perplexity which has important implications that will be discussed in-depth in this paper.

Although it has been dubbed a revolutionary breakthrough, Google's LaMDA has a predecessor called Meena, released in early 2020, which has a very similar architecture and training method. Meena has outperformed competing open-domain conversational systems at the time of release, and has made important contributions to the field of NLP. One significant contribution includes determining that perplexity can be used as an automatic evaluation of dialogue. In the past, automatic evaluators such as BLEU have been used, albeit with struggles when it comes to human-like dialogue.

Due to the limited technical information available on LaMDA, this report will examine Meena, discussing how it was developed by Google, its benefits and shortcomings, a comparison to other existing open-domain systems, and what we can expect in the coming years.

While Meena showed initial success at being both sensible and specific, LaMDA went a step further. With larger training data and parameter finetuning, LaMDA has improved on all four qualities. Although there is no data available to empirically prove its improvement to Meena, LaMDA is able to have more free flowing conversations and provide more meaningful and context-dependent yet unexpected and witty responses.

Meena Overview

Historically, developing open-domain chatbots has proven to be difficult, due to the chaotic quality of real-life conversation. Their answers either did not make sense or were very generic. For example, responding "I don't know," is a sensible answer to a question, but is not specific. Google's introduction of Meena has given the NLP field the building blocks to create a more human-like conversational system.

In this section we will detail Meena's methodology, how it was trained and its performance compared to existing open-domain chatbots, prior to our analysis of its performance and impact on NLP as a whole. First we will go over how Meena was trained:

Data Source

Meena is an end-to-end 2.6 billion parameter conversation model trained on a dataset of 40 billion words containing 341 GB of text with 4 billion different words. This is an incredible amount of data considering that the entirety of the English Wikipedia was 100 GB in 2015 and that GPT-2 was trained on 40GB of text. To obtain information they mined public domain social media multi-turn conversational data, which refers to data involving 2 or more speakers. An example of this would be a Twitter thread. They then trained the neural net to answer prompts by showing 7 interactions of conversation as input.

Certain data cleaning steps were taken to clean the data, such as removing conversations where:

- Usernames included the word "bot"
- Less than 70% of characters were alphabetic
- The message was repeated more than 100 times
- There was high overlap with parent's text (the first message)

The result of all these steps was 341GB of data based on 867 million context-response pairs. The data has been trained with a Evolved Transformer neural net, which has proven to be especially good at capturing context with its self-attention mechanism (to be discussed later).

Evaluation Criteria

Google created an evaluation metric known as the Sensibleness and Specificity Average (SSA) in order to measure Meena's performance relative to others. Humans in groups of 5 agreed on whether a chatbots' response was: a) sensible (an appropriate response for humans), and b) specific to the context. Specificity is important because generic responses were a big issue for previously developed chatbots. One reason this was common was because generic responses enabled chatbots to pass the Turing Test which in the initial years of NLP research was considered good performance by a chatbot. An analysis such as the Turing Test's "human likeness" has often been used elsewhere, but lacks objectivity and doesn't punish vague responses, both reasons that support SSA's usage.

SSA was chosen because it's more objective to select a response as being specific and/or sensible than something more subjective such as "human-like". However, it was ultimately realized that SSA correlates strongly with human-likeness and perplexity, which is an important discovery because it's an automatic evaluation metric. This will be expanded upon later on.

To see examples of both in practice:

1. Sensibleness but no Specificity (Bad Performance):
   Human: I really like playing soccer!
   Bot: Great. It sounds fun!

2. Sensibleness and Specificity (Good Performance):
   Human: I really like playing soccer!
   Bot: Great! My favorite team is Real Madrid, what is your favorite?

As mentioned above, humans voted on whether a response was Sensible and/or Specific. The SSA metric was based on two types of human evaluation:

1. Static Evaluation: a curated dataset of 1,477 multi-turn conversations
2. Interactive Evaluation: a human could chat about any topic of choice. Conversations started with "Hi!".

Performance Against Competing Open-Domain Chatbots

In order to compare the Meena's performance, Google decided to compare it against other well known open-domain chatbots: Mitsuku, Cleverbot, XiaoIce, and DialoGPT, as well as against human interactions. For the Interactive Evaluation component, 100 conversations were collected for each model, with the criteria of each one being that it must range between 14 turns and 28 turns (each output by a human or chatbot constitutes a turn). The result was 7 to 14 turns by the chatbot, and therefore 700 - 1,400 labeled responses.

To compare with human performance as well, 100 human-to-human conversations were also collected where they knew they were speaking with a human, rather than a chatbot. One potential area of bias here was that humans tend to say unusual things to chatbots in order to challenge them, compared to if they are speaking with a human.

Static Evaluation was also used with the purpose being to have a common benchmark that could be used to easily compare models. These 1,477 multi-turn conversations were taken from various sources such as essays and contest transcripts, and varied between 1-3 conversational turns. This raised the potential for bias, although researches justified this by highlighting strong correlation between Static Evaluation's scores with Interactive Evaluation's scores.

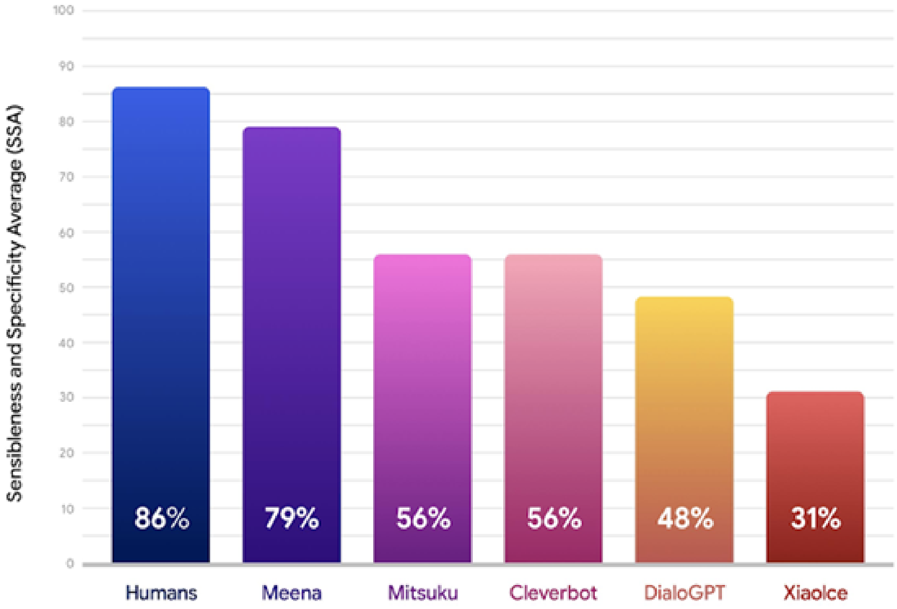
The Static Evaluation scores and Interactive Evaluation scores were averaged to get a final SSA score.

They accessed the other chatbots in various ways:

1. Cleverbot: used its API to integrate
2. DialoGPT: used the 762M parameter version, and used K=10 for top-K decoding (refers to the top-k most probable words)
3. Mitsuku: used the free web application
4. XiaoIce: no public API so they spoke on the public web app, and in Mandarin. They had trouble refreshing the bot which may have impacted results, as conversations may have been impacted by prior conversations.

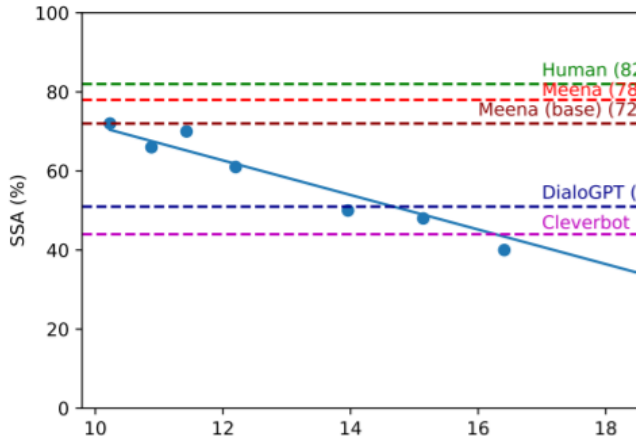
It's important to note here that Google fine-tuned their own model for SSA, whereas they didn't have full access to the other chatbots and did not finetune them for SSA scoring purposes. However, the large gap in performance scores likely supports that this wouldn't have been material in terms of performance.

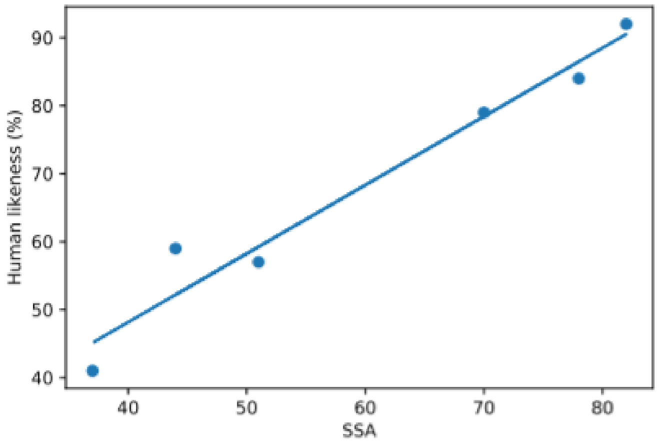
The results were as follows: Meena scored 79%, Mitsuku and Cleverbot 56%, DialoGPT 48%, and XiaoIce 31% and the average human score 86%.



It’s interesting to note that humans scored 86%, which was due to factors such as missed humour and lack of concentration. Meena originally scored 72% but was then fine-tuned to 79%, such as by eliminating the ability for common subsequences. If we consider 86% to be a human-level score, then Meena’s score of 79% is in fact quite close.

Notably, SSA correlated strongly with the perplexity score, which is an unsupervised way of measuring text generation quality. Note that a lower perplexity score signals higher ability to predict unseen words, and the lower the perplexity, the better the quality of generated text. The implication is that perplexity is an intrinsic evaluation method of dialogue and thus ideal for dialogue quickly (without human evaluator). Additionally, it correlated strongly with human-likeness, a crowd-sourced metric whereby humans evaluate to what degree the chatbot seems human-like. BLEU, a score that has been commonly used as an automatic metric for translations has shown to be relatively weak for dialogue evaluation, which is why it has not been considered as an evaluation metric here.

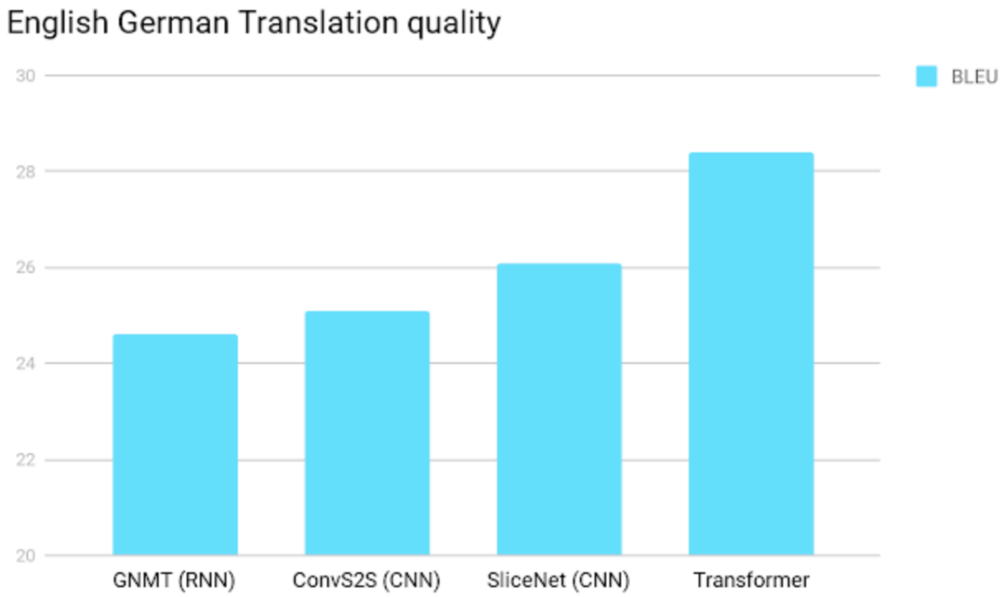




The Role of Transformer Neural Networks:

Meena relies on an Evolved Transformer (ET) seq2seq model. An ET is comparable to a regular Transformer, but rather than solely focusing on self-attention, it also considers wide convolutions through the use of Neural Architecture Search (NAS). The Transformer has had a huge impact in the field of NLP and is used in both LaMDA and Meena, and as such we will begin by discussing Transformers themselves.

Transformers are a neural net architecture originally developed and open-sourced by Google in 2018, which have now become state-of-the-art for many language models. Generative language models, such as GPT-3, and translation models such as BERT, use the Transformer architecture. Transformers have effectively replaced Recurrent and Convolutional Neural Nets, such as GNMT and SliceNet, as they have been shown to significantly improve the BLEU score (BiLingual Evaluation Understudy) for translation models, while reducing computational costs significantly.



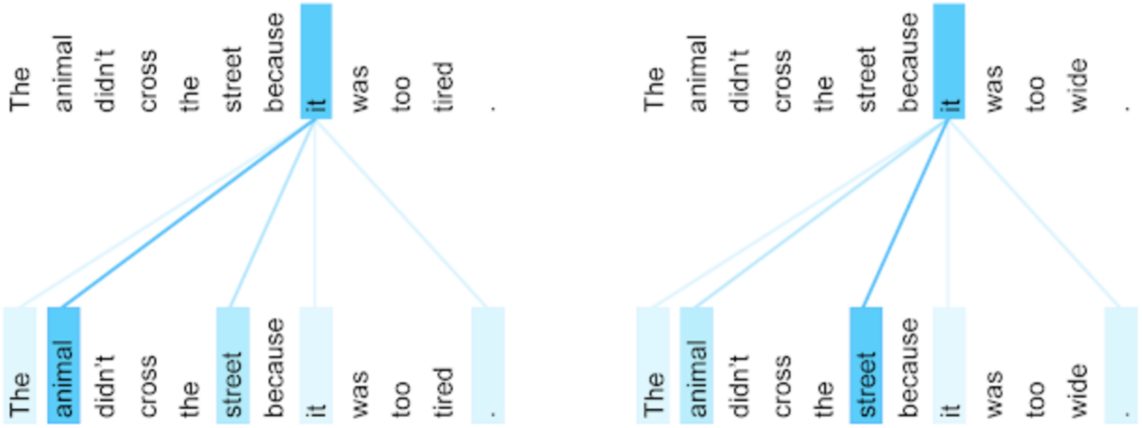
With Transformers being the standard neural net for translations, they have also performed incredibly for training chatbots, such as Meena.

The reason why neural nets perform so well for language models is due to their ability to aggregate information from surrounding words to determine the meaning of a given bit of language in context. This is necessary because a word, such as “bank”, can have two meanings depending on the context of the sentence. Bank can refer to the “bank” of the river or a financial institution.

Recurrent Neural Nets (RNNs) and Convolutional Neural Nets (CNNs) process the language sequentially in a left-to-right or right-to-left fashion. By reading one word at a time, it performs multiple steps to make decisions that depend on words far away from each other. By using the example of “bank” a RNN could only determine that “bank” is likely to refer to the bank of a river after reading each word between “bank” and “river” step by step. Naturally, the more steps the neural net has to compute, the harder it is for a recurrent network to learn how to make those decisions.

RNNs' sequential nature makes it more difficult to fully exploit new fast computing devices like TPUs and GPUs, which excel in parallel processing rather than sequential processing. CNNs are far less sequential than RNNs, but the number of steps necessary to incorporate information from distant regions of the input grows incredibly as the size of data grows. As such, there are computational issues as we scale.

In contrast, the Transformer only performs a small, constant number of steps. The Transformer applies a self-attention mechanism which is able to model relationships between all words in a sentence, independent of their position. In the earlier example of the “bank”, the self-attention mechanism of the Transformer is able to determine that the word “bank” refers to the shore of a river and not a financial institution in a single step. In order to predict a given word of a sentence - “bank” for example - the Transformer compares it to every other word in the sentence, which is used to compute an attention score for every other word in the sentence. These attention scores determine how much each of the other words should contribute to the meaning of “bank”. Continuing with the above example, “river” could receive a high attention score when computing a new meaning for “bank”. The attention scores are used as weights for a weighted average of the representation of all words which is fed into a fully-connected network to generate a new representation for “bank”, in order for the system to conclude that the sentence is talking about a river bank. An example of the usage of weights of the self-attention mechanism can be shown in the image below.



Various models were explored by the researchers at Google but ultimately the best performing Meena model was an Evolved Transformer (ET) model with 2.6B parameters that included 1 ET encoder block and 13 ET decoder blocks. The ET model, as mentioned prior, also uses self-attention but also incorporates Neural Architecture Search (NAS). NAS is an algorithmic approach to finding the optimal design for a specific neural network.

Transformer Decoder Tuning

Meena’s ET model has 13 decoder blocks. Historically, decoding frameworks such as adversarial networks have been used but result in increased complexity and reduced scalability. Scalability is quite important for a chatbot that can potentially be used by a vast sum of people down the road. For this reason, researchers at Google used the “Sample-and-Rank” method which was shown to give both high-quality and diverse responses.

In Sample-and-Rank, a parameter known as Temperature (T) is tuned. A high T leads to a response with common tokens (words), meaning more likely to be sensible but less likely to be specific. In contrast, a low T leads to contextually rare tokens which are therefore specific but more likely to not be sensible. Sample-and-Rank works by randomly sampling various potential responses, tuning the Temperature parameter, and selecting the candidate response with the highest score. By being accurate (low perplexity score) it can have a high Temperature score, resulting in a human-like answer that's both sensible and specific.

Meena's Contributions to NLP research

It's clear that Google's research with Meena has made significant contributions to NLP, showing that SSA can be used as an automatic metric for perplexity and human-likeness. Meena's shortcomings, however, must also be addressed to advance NLP research with subsequent results such as important use-case implementations and greater understanding of the human language.

It's important to highlight that being sensible and specific isn't everything when it comes to a chatbot. Meena lacked in regards to other potentially important characteristics that make humans human, such as being interesting, empathetic, humorous, or sad. While it outperformed other open-domain bots in SSA, some of those bots excelled in other areas, such as XiaoIce which notably was more empathetic and had a stronger personality. These characteristics can be important for various use-cases, such as being a digital companion in elderly homes. Additionally, it struggled with free-flowing conversations and made errors such as contradicting itself.

Ethical nuances must also be taken into consideration. Chatbots responses are unsupervised, thus there are some ethical issues, biases and safety concerns. Despite not explicitly stating which biases are inherent in Meena, Google mentioned that it will not release an external research demo due to challenges that it has faced regarding this problem. The consequences of unethical AI can be severe. The responses of previous chatbots from other providers have been unintentionally racist or offensive. For example in a Jeopardy competition the participants had to provide an answer that only contained a 4-letter word, Watsons IBM's chatbots answer was "F*ck". Clearly, it could find a good response but was unable to comprehend that it was not appropriate given the context of being on a game show. Tackling safety and bias in the models is as mentioned above are some of the key focus areas for Google AI and other leading competitors.

Advancing the research on open-domain chatbots is critical to the field of NLP, as their discoveries can in-turn be used elsewhere, such as better closed-domain chatbots and audio generation.

LaMDA seems to have made significant advances upon Meena in chatbot technology, clearly making machine interactions more human-like and intuitive. As artificial intelligence and computational power capabilities continue to improve, we will continue to see further advancements with likely very surprising impacts and results. Similar to how LaMDA was built upon Meena, future research will likely build upon LaMDA as the world moves incrementally closer to a seamless interaction with machines.

References:

Google AI Blog: Towards a Conversational Agent that Can Chat About…Anything (googleblog.com)
Just how big a deal is Google's new Meena chatbot model? | VentureBeat
A Breakthrough in Chatbot: Review of Meena | by AI Frontiers | AI Frontiers | Medium
Artificial intelligence: Does another huge language model prove anything? – TechTalks (bdtechtalks.com)
Book: How Smart Machines Think
Personalizing Open Domain Chatbots: How OLPORTAL Holds the Key | by OLPORTAL.ai | Medium
[2001.09977v3] Towards a Human-like Open-Domain Chatbot (arxiv.org)
https://medium.com/analytics-vidhya/chat-about-anything-with-human-like-open-domain-chatbot-7649408fe279
https://towardsdatascience.com/deep-learning-how-to-build-an-emotional-chatbot-part-2-the-dialogue-system-4932afe6545c
https://www.kdnuggets.com/2020/02/inside-machine-learning-google-build-meena-chatbot.html
https://www.infoq.com/news/2021/01/google-microsoft-superhuman/

Structure Discussed:

1. Introduction
    a. Overview of our report, quick mention of importance of Meena in NLP research, etc
    b. Brief history of chatbots, open-domain vs closed-domain, why chatbots are important
2. Meena overview
    a. Methodology, evaluation criteria, what they did to fine-tune it (explained in Notes, like when they took it from 72% SSA to 79%)
    b. Success attributed to:
        i. Large dataset
        ii. Solid evaluation criterion
        iii. Neural network architecture generated by NAS (neural architecture search)
    c. Brief description of Transformers (personally I don't consider it critical to our analysis of Meena but yes let's discuss it a bit)
3. Meena Analysis
    a.

Introduction

- History of chatbots
- Why research in this area is important
- What we will discuss in our paper