

# Towards a Human-like Open-Domain Chatbot

- 저자:
  - Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, Quoc V. Le (**Google Research, Brain Team**)
- 발표:
  - Presenter: 윤주성
  - Date: 200206

## Who is an Author?

- 구글스칼라나 다른 곳에 딱히 프로필이 없음
- 그의 행적은 트위터에..  
<https://twitter.com/xpearhead>
- 미디엄도.. <https://medium.com/@dmail07>



## 느낀점

- 일단 논문이 꽤 길다
- 모델쪽보단 automatic evaluation metric을 제안했다는것에 은근 더 중점을 맞추는 느낌
- 모델쪽 얘기는 Evolved Transformer논문을 더 봐야할듯
- 뭐랄까.. 설명이 많고 장황한 논문이다. 새로운 개념을 정의하는게 많은 논문임. 제안하는 개념이 필요한 이유등을 주로 설명함.
- Metric + large scale + tip이 본 논문의 주요 contribution인듯 modeling적인 부분은 별로 기술되어있지 않음

## Abstract

- Meena라는 이름을 가진, multi-turn open-domain chatbot을 제안함
- public domain social media conversation으로부터 데이터를 정제해서 end-to-end로 학습시킴 (~~데이터 공수가 꽤 중요했을듯~~)
- 2.6B params의 neural net으로 이루어진 모델은 단순히 next token의 perplexity를 최소화시키는 방법으로 학습함
- Sensibleness and Specifity Average (SSA)라는 Human evaluation metric도 제안함 (human-like multi-turn conversation의 key element를 평가)
- 실험 결과로 perplexity와 SSA간에는 strong correlation이 있는 것으로 나타남
  - best perplexity를 갖는 모델은 SSA도 높았음
  - The fact that the best perplexity end-to-end trained Meena scores high on SSA (72% on multi-turn evaluation) suggests that a human-level SSA of 86% is potentially within reach if we can better optimize perplexity
- full version of Meena (filtering mechanism + tuned decoding)의 경우 79% SSA점수를 받았음 (기존에 존재하는 챗봇들보다 23% 높았음)

# 1. Introduction

- 자연어로 자유롭게 대화하는 능력은 human intelligence의 증표 같은 것이며 진정한 인공지능에게 요구되는 것임
- closed-domain 챗봇은 특정 태스크를 위해 키워드나 인텐트등에 답변하지만 open-domain 챗봇의 경우엔 any topic에 대해서도 답변이 가능해야함
- 기존 연구에서 MILABOT (Serban et al., 2017), Xiaolce (Zhou et al., 2018)<sup>1</sup>, Gunrock (Chen et al., 2018), Mitsuku (Worswick, 2018)<sup>2</sup> and Cleverbot<sup>3</sup> (by Rollo Carpenter) 등이 제안되었지만, complex frameworks (dialog managers with knowledge-based, retrieval-based, or rule-based systems)에 의존적이었음 (~~어떻게 보면 당연한거같기도 한데 여기서의 지적함~~)
- End-to-End NeuralNet 접근방법도 있었지만 하나의 학습 모델로 평이한 결과를 냄

# 1. Introduction

- 많은 연구에도 불구하고 open-domain chatbot은 여전히 약점들을 갖고 있음
  - open-ended input에 말이 안되는 답변을 하곤함
  - 혹은 모호하거나 일반적인 답변을 함
- 본 연구에서는 generative chatbot인 Meena를 제안함
  - 40B words mined and filtered from public domain social media conversation 으로 end-to-end 학습함
  - end-to-end 접근방법을 한계까지 밀어붙임
  - large scale low-perplexity model이 대화 잘하는지 확인함
- Meena에서는 Evolved Transformer (So et al., 2019) 기반의 seq2seq model을 사용함
  - 모델은 multi-turn conversation 으로 학습함
  - input sequence 형태는 all turn of the context 임 (최대 7개까지 사용) (핑퐁팀의 접근과 비슷한듯)
  - output sequence 형태는 response임
  - best model은 2.6B params을 사용하고 test perplexity로 10.2를 기록함

# 1. Introduction

- Meena와 다른 챗봇의 quality를 평가하기 위해 Sensibleness and Specificity Average (SSA)를 제안함
  - 모델의 모든 답변에 대해 사람이 label을 달 때, 저 두가지 기준을 사용함
  - sensible은 context 안에서 말이 되는지를 판단함
    - 사람이 만든 대화에서는 97%정도가 이 기준에 부합했음
    - 하지만 이것만 가지고 판단할 경우 답변이 다소 모호하거나 지루할 수 있음(vague and boring)
    - 예를 들면, "**I don't know**" 답변이 계속 나온다면 하는 문제가 발생함
    - 그렇기 때문에 또 다른 평가 기준이 필요함
  - specificity는 주어진 context안에서 답변이 얼마나 구체적인지를 평가함
- Meena, humans 그리고 다른 오픈도메인 챗봇을 SSA metric으로 비교하기 위해 2가지 타입의 human evaluation을 사용함
  - static:
    - 1,477 multi-turn conversations을 큐레이션해서 데이터셋을 만듦
  - interactive:
    - 사람이 시스템에 원하는건 무엇이든지 입력함
  - surprised but pleased, static, interactive evaluation에서 모두 SSA metric이 Meena's perplexity와 strong correlation이 있음을 발견했음 ( In other words, the better that Meena fit its training data, the more sensible and specific its chat responses became. )
  - 어찌보면 당연해보이는게 왜 놀랍냐면, 최근 연구들에서는 human evaluation과 BLEU같은 automatic metrics이 poor correlation을 갖는게 밝혀졌기 때문임 (~~사실 BLEU를 생성 모델의 평가지표로 쓰는거 자체가 말이 안되긴 함~~)

# 1. Introduction

- best end-to-end learned model은 avg 72% SSA를 기록했고 full version of Meena (+filtering mechanism and tuned decoding)는 79%의 SSA를 기록함 (사람 간의 대화 점수는 avg 86%임, 사람의 경우 sensibleness가 매우 높고, specificity는 눈에 띄게 낮은 경향이 있었음)
- 평가방법에 단점이 있다면, static eval의 경우 데이터셋이 제한되어있으므로 human conversations을 모두 커버한다 할 수 없음
- 그럼에도 불구하고 SSA score와 perplexity간에 발견한 correlation을 통해 perplexity를 개선하는게 human-like chatbot에 도움이 된 다걸 발견했다는데 의의가 있음
- 본 논문의 contribution은
  - proposing a simple human evaluation metric for multi-turn open-domain chatbots
  - showing evidence that perplexity is an automatic metric that correlates with human judgment
  - demonstrating that an end-to-end neural model with sufficiently low perplexity can surpass the sensibleness and specificity of existing chatbots that rely on complex, handcrafted frameworks developed over many years



## 2. Evaluating chatbots

- 챗봇과 NLG를 평가하는건 잘알려진 challenge임 (Liu et al., 2016; Lowe et al., 2017; Novikova et al., 2017; Hashimoto et al., 2019)
- 본 논문에서는 사람과 같은 답변을 하는지를 평가하기위한 human evaluation metric을 제안함
- 2가지 setup인 static (fixed set of multi-turn contexts to generate responses), interactive(chat freely with chatbots) 으로 나누어서 진행함

## 2.1 Measuring Human Likeness

- 주어진 context에서 말이 되는지를 평가 ( given the context, makes sense )
- Sensibleness
  - 평가 요소
    - common sense
    - logical coherence
    - consistency
  - 약점
    - However, being sensible is not enough. A generic response (e.g., I don't know) can be sensible, but it is also boring and unspecific
    - GenericBot(question에는 항상 "I don't know"를, statement에는 항상 "ok" 시전)을 만들어서 평가해보니 static evaluation에서 DialoGPT (62%) 보다 높은 70%의 Sensible 점수를 받음 (DialoGPT 대답이 더 human-like 했음에도!)
    - 이러한 문제를 개선하기 위해 답변이 sensible로 label되면, crowd worker에게 이 답변이 context에 맞게 충분히 specific한지를 평가함
- Specificity
  - 일단 sensible label이 된 상태여야 specificity label 할지를 고려함
  - 반례:
    - A says: "I love tennis,"
    - B responds: "That's nice,"
    - label: "not specific"
  - 올바른 예:
    - A says: "I love tennis,"
    - B responds: "Me too, I can't get enough of Roger Federer!"
    - label: "specific"
  - GenericBot의 경우 어떠한 답변도 "specific" label을 받지 못했지만 DialoGPT의 경우 39%정도 "specific" label을 받음

## 2.1 Measuring Human Likeness

- 평가할때 들어가는 주관의 정도는 crowd worker의 agreement로 정량화 할수 있음 ( The degree of subjectivity is somewhat quantified in the crowd worker agreement )
- 모든 모델 성능평가에 대한 crowd worker의 consistency를 측정하기 위해 agreement와 Krippendorff's alpha (Krippendorff, 2011)를 사용함 ([https://en.wikipedia.org/wiki/Krippendorff's\\_alpha](https://en.wikipedia.org/wiki/Krippendorff's_alpha))
- Sensibleness와 Specificity를 하나의 metric으로 사용하기 위해 간단히 두 값을 평균냈고 이를 SSA (Sensibleness and Specificity Average)라 표현
- 점수:
  - GenericBot: 35%
  - DialoGPT: 51%

Metric	Agreement (%)	Krippendorff's alpha
Sensibleness	$76 \pm 3$	$0.42 \pm 0.03$
Specificity	$66 \pm 2$	$0.30 \pm 0.05$

Table 1: The average and standard deviation of crowd worker agreement across static evaluations of Meena models. Each static evaluation consisted of 1,477 (*context, response*) pairs, each labeled by 5 crowd workers.

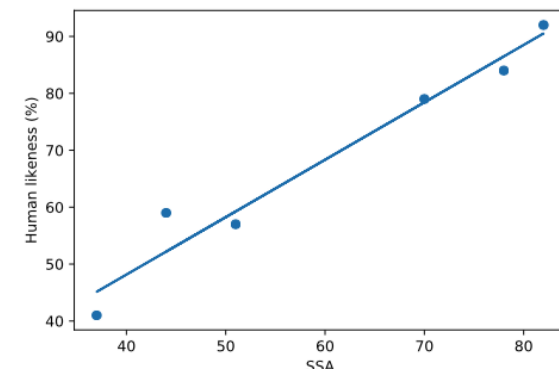


Figure 2: SSA vs human likeness. Each point is a different chatbot, except for the top right one, which is human. A regression line is plotted, for which the coefficient of determination ( $R^2$ ) is 0.96. The SSA values were collected using static evaluation mode (Section 2.2). The human likeness evaluation was also conducted in static evaluation mode. Instead of judging sensibleness or specificity, however, we asked crowd workers to judge whether a given response was “human-like”, or in other words, looked like a response that a human might give in the provided context.

## 2.2 Static Evaluation

- 모델 비교를 편하게 하기 위한 common benchmark가 필요했음
- 1,477 conversational contexts (with 1~3 conversation turn)
- 위의 조건을 만족하는 대화 데이터셋을 Mini-Turing Benchmark (MTB)라 정함
- turn 개수에 따른 데이터 수
  - single-turn context (e.g., "How are you?"): 315
  - two-turn: 500
  - three-turn: 662
- MTB에는 personality question도 포함됨 (e.g. "Do you like cats?")
- 이때 personality는 일관성이 있어야함
  - For example, the context "A: Do you like movies?; B: Yeah. I like sci-fi mostly; A: Really? Which is your favorite?" expects a consistent response such as I love Back to the Future. On the other hand, a response like I don't like movies would be a contradiction
- (context, response) pair에 대한 결과를 crowd workers에 보여준뒤 sensible and specific을 평가함
- 이때 static이라는 단어가 왜 붙냐면 contexts가 고정되어 있기 때문임

## 2.3 Interactive Evaluation

- static eval이 모델 비교하기엔 편하나, dataset에 따라 biased 한 경향이 있음
- crowd workers가 1:1로 chatbot에게 아무거나 물어보고 싶은걸 물어보는 평가 모드로 하나 더 추가함
- 대화는 최소 14턴이 요구되고 (7턴은 챗봇), 최대로는 28턴까지 실험함
- 각 모델당 100개의 conversations을 수집함 (즉 최소, 700개의 labeled turns이 모델당 있는 것임)

## 2.4 Estimate of Human Performance

- internal company volunteers의 도움을 받아 100개의 human-human conversations을 수집했음
- SSA를 위한 labeling은 위의 volunteers와는 상관없는 5명의 crowd workers의 majority voting으로 매 human turn마다 매겨짐

## 2.5 Evaluation of Cleverbot and DialoGPT

- Cleverbot은 API 사용했음
- DialoGPT는 762M params짜리 공개된 모델을 사용함 (345M짜리 모델이 single-turn에서 가장 좋다고 알려져서 먼저 사용했었는데 762M 모델에 비해 multi-turn에서는 성능이 현저히 성능이 나빠서 762M 사용)
- DialoGPT authors가 decoding script를 공개하지 않아서, top-K (K=10) decoding을 적용했음 (코드는 huggingface의 Wolf가 공개한 구현체 씬)
- 위의 두 모델도 Meena와 같이 crowd sourcing으로 평가됨

## 2.6 Evaluation of Mitsuku and Xiaolce

- Mitsuku는 webapp을 사용해야했고 Xiaolce는 public API 없어서 interactive evaluation을 수행
- 자원봉사자들이 Mitsuku는 100개의 대화를, Xiaolce에서는 119개의 대화를 수집함



## 2.7 Automatic Evaluation

- quick research iterations을 위해, perplexity에 집중했음
- 기존의 평가방법과 달리, perplexity가 automatic metric가 됨
- A seq2seq model outputs a probability distribution over possible next response tokens
- Perplexity measures how well the model predicts the test set data; in other words, how accurately it anticipates what people will say next. When interpreting perplexity scores, bear in mind that lower is better and that the theoretical minimum is one

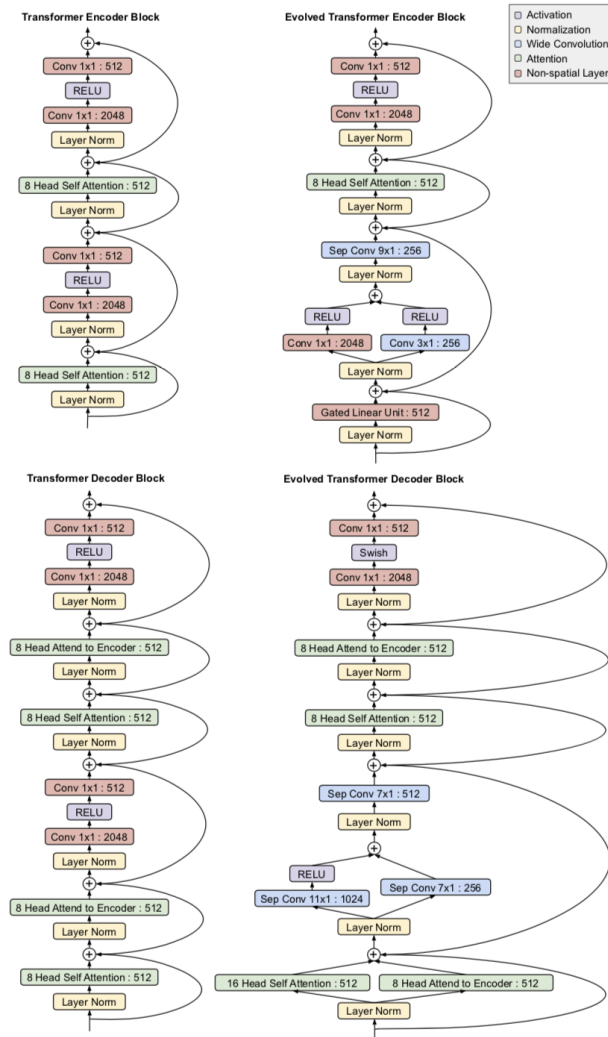
### 3. Meena chatbot

- 기존의 e2e dialog models은 크게 두가지 종류로 나뉘었음
  - (1) complex models with human-designed components
  - (2) large neural network models (known as end-to-end models)
- 핵심 질문!! (open research question)
  - in order to reach a point where a model can carry out high-quality, multi-turn conversations with humans, could we simply take an end-to-end model and make it bigger—by adding more training data and increasing its parameter count—or is it necessary to combine such a model with other components?
- 본 논문이 제안하는 Meena Model (the largest end-to-end model)이 humanlike chat responses 를 open domain에서 할 수 있다는 걸 보여줌으로써 위의 open research question에 답변이 가능할 것이라고 주장

## 3.1 Training Data

- Data는 public domain social media conversations을 정제해서 만듦
- source data는 multiple speakers를 포함한 message trees 형태임
- 첫번째 메시지를 root로 삼고, 거기에 대한 답변이 child 노드가 됨
- 이런식으로 하면 대화할때 각 메시지의 conversation turn을 알 수 있음
- 대화속의 각 턴을 답변으로 하고 이전 턴들을 context로 하면 (*context, response*) pair training dataset을 만들 수 있음
- generation quality를 높이기 위해, 다음과 같은 조건을 만족하면 메시지를 삭제함
  - a. subword의 개수가 2개보다 적거나 128개보다 많은 경우
  - b. alphabetic characters의 percentage가 70% 이하인 경우
  - c. message가 URL을 갖고 있는 경우
  - d. author's username에 "bot"이 포함된 경우
  - e. 메시지가 100번 이상 반복된 경우
  - f. parent's text와 high n-gram이 겹치는 경우
  - g. commercial text classifier가 메시지가 안전하지 않거나 공격적이라고 분류한 경우
- 추가로 메시지 내에서 parent's text를 인용한건 따로 제거했음
- 메시지가 제거된 경우 그 메시지 기준 sub-tree는 다 drop시킴
- 필터링 후에 남은 (*context, response*) pair의 개수: 867M
- tokenization: BPE with sentencepiece
- vocab size: 8K BPE subwords (이전 실험들에서 이정도 사이즈로도 답변을 생성하기에 충분하다는걸 확인함)
- Final Meena dataset 크기: 341GB of text (40B words) (GPT-2의 경우 40GB of internet text(8 million web pages) 사용)

## 3.2 Model Architecture



**Figure 3. Transformer and Evolved Transformer architecture cells.** The four most notable aspects of the found architecture are the use of 1) wide depth-wise separable convolutions, 2) Gated Linear Units (Dauphin et al., 2017), 3) branching structures and 4) swish activations (Ramachandran et al., 2017). Both the ET encoder and decoder independently developed a branched lower portion of wide convolutions. Also in both cases, the latter portion is almost identical to the Transformer.

## 3.2 Model Architecture

- Meena model에서 가장 성능이 좋았던건, Evolved Transformer (ET) (So et al., 2019) seq2seq model임 (with 2.6B parameters)
- 1 ET encoder block + 13 ET decoder blocks 사용
- ET는 transformer를 기반으로한 evolutionary NAS architecture임
- largest ET 모델은 10.2 perplexity를 기록했고 largest vanilla transformer는 10.7 perplexity를 기록함 (모두 738k step으로 고정했을 때)
- vanilla Transformer는 32 decoder layer를 사용했고 나머지 hyper params는 동일함 (transformer 논문에선 6개, 버트에선 base가 12개 large가 24개임) (이러한 이유 때문에 이렇게 맞춰준듯 `An Evolved Transformer block is about twice as deep as a Transformer layer`)
- 다른 모델과 비교하자면, extra-large GPT-2는 1.5B params을 갖고, LM 기반임(decoder only), DialoGPT의 경우엔 대화모델이고 762M params을 가짐
- Meena's hyper Params
  - hidden size: 2,560
  - attention head: 32
  - share embeddings across the encoder and decoder
  - encoder, decoder maxlen: 128 tokens (256 combined)
- optim 방법: manual coordinate-descent search

### 3.3 Training Details

- TPU-v3 Pod (2,048 TPU cores)에서 30일간 (~~또 이렇게 돈을 태웠..~~) 학습함
- dataset: 40B words (or 61B BPE tokens)
- 2.6B-param model이 61B-token dataset에도 overfit됨 (모델 capa가 엄청 크다는걸 보여줌)
- 이러한 이유로 add a small amount of 0.1 attention and feed-forward layer dropout (~~0.1 attention을 더했다는게 뭐 자..~~)
- 메모리를 아끼기 위해, Adafactor optimizer (Shazeer and Stern, 2018)를 사용했음
  - init lr: 0.01 (0~10k steps)
  - decaying lr with the inverse square root of number of steps (10k step 이후 적용)
- Tensor2Tensor codebase 사용함
- TPU 셋팅
  - TPU-v3의 core는 16GB of high-bandwidth memory를 가짐
  - 메모리 사용량 최대한으로 각 코어당 8 training examples를 할당시킴
  - 그 결과 1 step에 1초 정도 걸림
  - full TPU-v3 Pod에서 4M tokens를 1초에 학습시킬 수 있음
  - 학습이 다 끝났을 때 모델은 164 epoch을 돌고 10T tokens을 봄(repeated tokens 포함)

## 3.4 Decoding

- generic and bland response를 생성하는건 neural conversational models에게 항상 문제였음
- 이러한 문제를 해결하기위한 common approach는 더 좋은 decoding algorithm을 사용하는 것 ( reranking, conditioning, adversarial learning, variational autoencoding 등등)
- 제안 모델의 경우엔 충분히 low perplexity를 가지기 때문에 a simple sample-and-rank decoding strategy로도 diverse and high-quality responses 를 만들어낼 수 있음
  - Sample-and-rank
    - a. Sample N independent candidate response (using plain **random sampling** with temperature  $T$ )
    - b. candidate response중에서 highest probability를 갖는 걸 final output으로 선택
- Temperature  $T$ 는 hyper param이고 the next token에 대한 확률분포  $p_{\{i\}}$ 를 regulate함
- Hinton et al. (2015)이 했던것 처럼 logits  $z_{\{i\}}$ 를 softmax 계산하기 전에  $T$ 나눴음

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

- $T$ 값에 따른 변화
  - $T = 1$  이면, unmodified distribution임
  - $T$ 가 커지면, contextually rare tokens 더 볼 수 있음 (relevant entity names)
  - $T$ 가 작아지면, common words가 더 많이 나옴, 안정적이거나 specific은 떨어짐 (예: 관사, 전치사)

### 3.4 Decoding

- “Why do you like the ocean?”이라는 질문이 input으로 있을 때 결과
  - beam search는 반복적인 답변, uninteresting 답변이 많지만 sample-and-rank의 경우 다양한 답변 및 context-rich 답변이 가능
  - 예시

Responses for “Why do you like the ocean?”	Score
1. I love the beach. I like that it doesn't have anything in it. Just open water.	-12.98
2. It's dark and mysterious. I discovered it when I was really young and it was a dream of mine since. The ocean is full of wonders, mysteries and great things to discover. I love snorkeling. I moved to Hawaii a year ago and I'm going to college for marine biology. So there's that.	-14.03
3. The sea sounds cool and it's really satisfying to swim in.	-14.62
4. I dunno. It's beautiful, I guess. It reflects my upbringing. I reckon.	-14.87
5. Well I'm from the pacific northwest and always somehow end up back there. I've been to Hawaii but I'm not one for the beach, it's too hot and the air is so dry.	-15.36
6. The beauty of it is incredible, and swimming in it and being able to see huge waves sweeping by is just amazing. I was born in the desert so it's really the only place in the world I've seen with flowing water around me	-17.07
7. It's beautiful, maybe? I just wish I lived closer to it!	-17.14
8. It's pretty and soothing.	-17.19
9. Because I love the smell. I love walking when I can through the beach. I live in California so I basically go to the ocean almost every weekend.	-17.38
10. The ocean is pretty cool, I suppose. I like the various types of fish.	-17.90

Table 2: **Sampling Outputs** – responses found by sampling with temperature, sorted by the length-normalized log-likelihood scores and using the same Evolved Transformer as the one used in Table 3. The scores are calculated as  $\frac{\log P}{T}$ , where  $P$  is the likelihood of the response and  $T$  is the number of tokens.

Responses for “Why do you like the ocean?”	Score
1. Why don't you like the ocean?	-1.70
2. I don't like the ocean. I love the ocean.	-2.66
3. I don't like the ocean. I love it.	-2.78
4. I don't like the ocean, I love the ocean.	-2.94
5. I don't like the ocean, I love it.	-2.94
6. Why don't you like the ocean? :P	-2.95
7. I don't like the ocean, I love it!	-3.15
8. I don't like the ocean. I love the ocean!	-3.20
9. Why don't you like the ocean? It's beautiful.	-3.26
10. I don't like the ocean. I love the ocean.	-3.31
There's a difference.	

Table 3: **Beam Search Outputs** – top responses generated by beam-search decoding and the corresponding length-normalized log-likelihood scores. We use an Evolved Transformer with perplexity 10.2 and vocabulary size of 8K.



### 3.4 Decoding

- Key point!
  - with low perplexity so samples can be taken at high temperature to produce human-like content
- sample-and-rank에서  $N = 20$ ,  $T = 0.88$ 로 셋팅
- Figure 1을 보면, decoding strategy를 고정했을때 perplexity를 개선하면 SSA가 높아지는걸 볼 수 있음

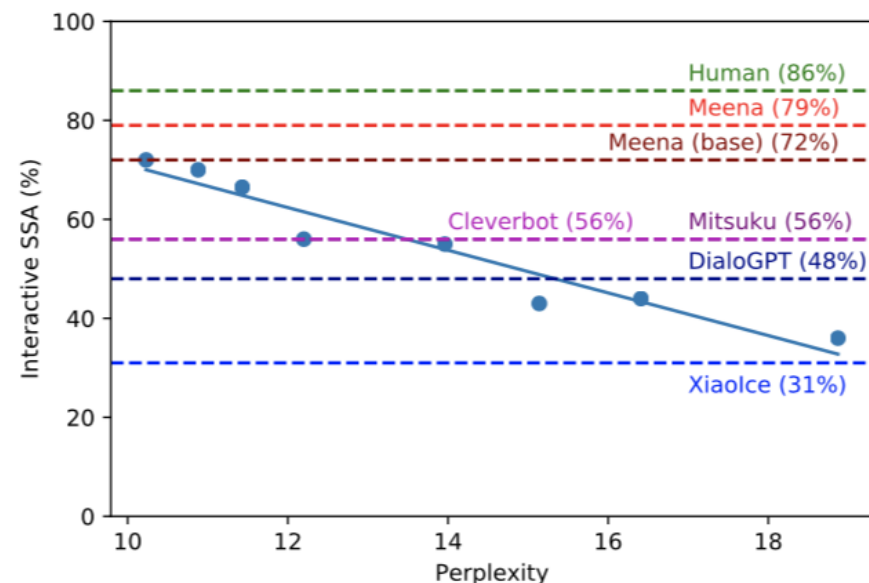


Figure 1: Interactive SSA vs Perplexity. Each point is a different version of the Meena model. A regression line is plotted, for which the coefficient of determination ( $R^2$ ) is 0.93, an indication of strong correlation between perplexity and the human evaluation metric (SSA). The dotted lines show the SSA performance of other chatbots, humans (86%), the best end-to-end trained Meena model (72%), and the full version of Meena which incorporates a filtering mechanism and tuned decoding (Section 5) and scores 79%. Mitsuku and Cleverbot scored the same on overall SSA, but Mitsuku displayed higher sensibleness, whereas Cleverbot had higher specificity. See Sections 2.5, 2.6, and 4.3 for more details on how we performed these comparisons and how to interpret the results.

## 3.5 Sample conversations

- Human과 Meena간의 대화 샘플 (~~체리피킹이라고 저자가 직접 써놓았음~~)
- 대화 생성때는 sample-and-rank를 사용했음
- Meena가 open-domain에서 대화를 그럭저럭 잘하지만 "Is it indoors or outdoors?"라고 묻는 부분을 보면 not sensible한 부분도 있음을 확인할 수 있음

## 3.5 Sample conversations

- 첫번째 예제:

---

Conversation A	
<b>Human:</b>	<i>say you're in an empty giant office and you see a black bird flying indoors. what do you do?</i>
<b>Meena:</b>	<i>What kind of bird? Is it friendly? How large? How big is the office? Is it indoors or outdoors?</i>
<b>Human:</b>	<i>well it's a predator bird and it's gonna eat you</i>
<b>Meena:</b>	<i>I would grab my phone and call 911.</i>

---

Meena handles arbitrary open-domain input. However, asking “Is it indoors or outdoors” was not sensible.

### 3.5 Sample conversations

- 두번째 예시: context 를 사용해서 대화하기도함 (내용이 실제로 맞음..)

Conversation B	
<b>Human:</b>	<i>Hi!</i>
<b>Meena:</b>	<i>Hi! How are you this fine evening?</i>
<b>Human:</b>	<i>I'm doing well. it's morning my time!</i>
<b>Meena:</b>	<i>Good morning! What are your plans for today?</i>
<b>Human:</b>	<i>Get my work done then watch a movie at home in the night</i>
<b>Meena:</b>	<i>Sounds fun! What movie?</i>
<b>Human:</b>	<i>I haven't decided yet, perhaps an Oscar-nominated one. any suggestion?</i>
<b>Meena:</b>	<i>I'm not too familiar with recent Oscar winners, but I do know I wanted to see The Grand Budapest Hotel.</i>
<b>Human:</b>	<i>ah, tell me about that movie!</i>
<b>Meena:</b>	<i>It's from the same director and writer as Moonrise Kingdom. Wes Anderson is pretty well known for making funny, colorful films. I've heard it's pretty good.</i>
<b>Human:</b>	<i>ha! I've searched on the internet and both movies are very colorful!</i>
<b>Meena:</b>	<i>I'm looking forward to it. I love colorful movies.</i>

Meena chats naturally about movies. The Grand Budapest Hotel and Moonrise Kingdom are both indeed colorful movies, directed by Wes Anderson.

### 3.5 Sample conversations

- 세번째 예시: 철학얘기하는 챗봇

---

Conversation C	
<b>Human:</b>	<i>Haha, I think mathematics has a presence beyond human existence.</i>
<b>Meena:</b>	<i>So does philosophy.</i>
<b>Human:</b>	<i>Can you state with examples how this is true.</i>
<b>Meena:</b>	<i>Can you state how it isn't?</i>
<b>Human:</b>	<i>Without human existence there will be no one left to comprehend whats right whats wrong, and thus philosophy wi [sic]</i>
<b>Meena:</b>	<i>That's not the definition of philosophy.</i>
<b>Human:</b>	<i>Without human existence there will be no one left to comprehend whats right whats wrong, and thus philosophy will not be needed. How do you define philosophy?</i>
<b>Meena:</b>	<i>The study of general knowledge or the fundamental nature of things.</i>

---

Meena **tries to define philosophy** in a way that is consistent with its previous statement.

### 3.5 Sample conversations

- 네번째 예시: 멀티턴 환경에서 농담하는 챗봇
- 자세한 대화 데이터셋은 Github 참고:  
<https://github.com/google-research/google-research/tree/master/meena/>

---

Conversation D	
<b>Human:</b>	<i>I like how cute cows are</i>
<b>Meena:</b>	<i>Ah, interesting. You like how they are cute, right?</i>
<b>Human:</b>	<i>yeah, and they're really smart</i>
<b>Meena:</b>	<i>I heard they go to college</i>
<b>Human:</b>	<i>cows go to college?</i>
<b>Meena:</b>	<i>I heard that a cow went to Harvard.</i>
<b>Human:</b>	<i>what did the cow study?</i>
<b>Meena:</b>	<i>Bovine sciences.</i>
<b>Human:</b>	<i>do horses go to Harvard?</i>
<b>Meena:</b>	<i>Horses go to Hayvard.</i>
<b>Human:</b>	<i>that's a pretty good joke, I feel like you led me into it</i>
<b>Meena:</b>	<i>You were trying to steer it elsewhere, I can see it.</i>

---

Meena executes a multi-turn joke in an open-domain setting. We were unable to find this in the data.

## 4. Results

- test perplexity와 human evaluation metric, SSA 간의 correlation에 대해 다루려함

## 4.1 SSA-perplexity correlation

- the number of layers, attention heads, total training steps 등을 바꾸기도 하고 ET쓸지 regular Transformer 쓸지, hard labels 쓸지 soft labels 등 고민하며 실험했음
- static eval에서 correlation이 거의 선형으로 보였지만 lower values of perplexity에서도 잘되는지 확인하고 싶어서 interactive eval도 수행했고 잘나온걸 확인 (dataset으로 인한 bias 문제는 없다고 주장할 수 있게 됨)

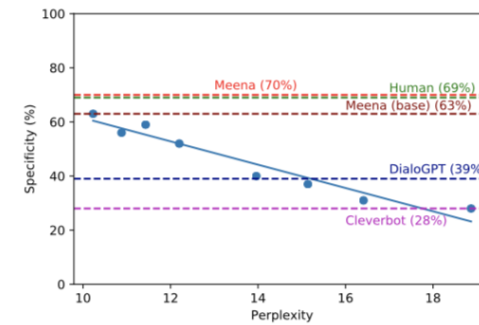
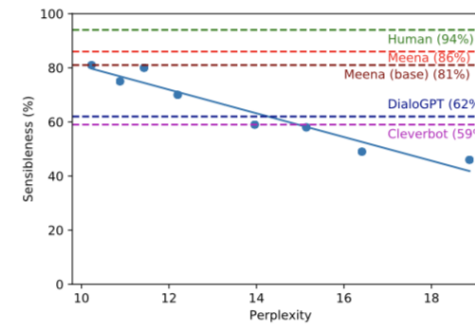
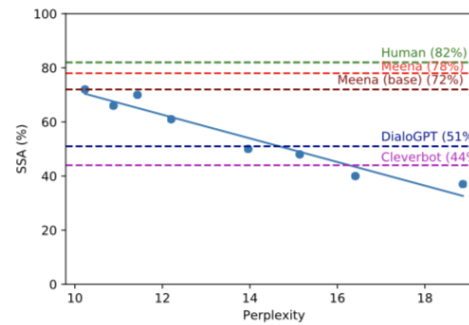


Figure 5: Static evaluation results.



## 4.1 SSA-perplexity correlation

- the number of layers, attention heads, total training steps 등을 바꾸기도 하고 ET 쓸지 regular Transformer 쓸지, hard labels 쓸지 soft labels 등등 고민하며 실험했음
- static eval에서 correlation이 거의 선형으로 보였지만 lower values of perplexity에서도 잘되는지 확인하고 싶어서 interactive eval도 수행했고 잘나온걸 확인 (dataset으로 인한 bias 문제는 없다고 주장할 수 있게 됨)
- the lowest perplexity model was evaluated 7 times with static evaluations and also 7 times with interactive evaluations.

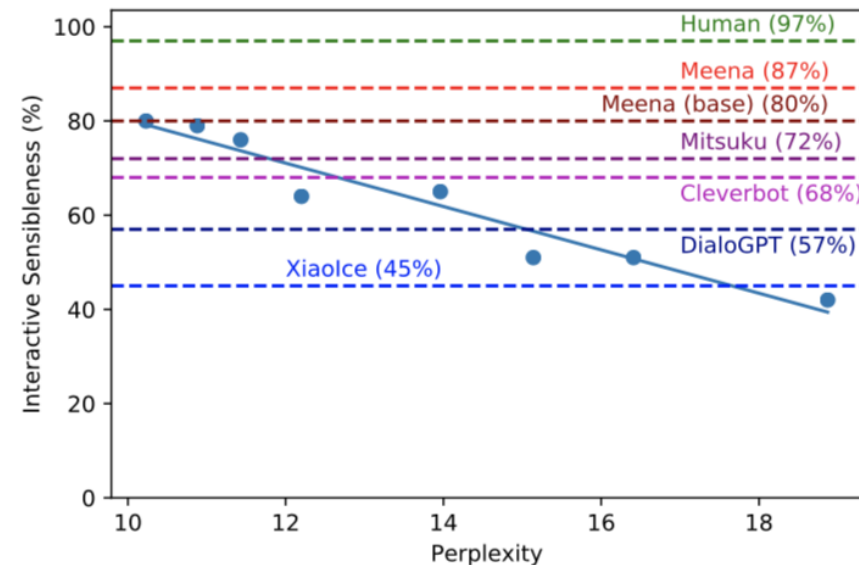


Figure 3: Interactive **sensibleness** vs perplexity.

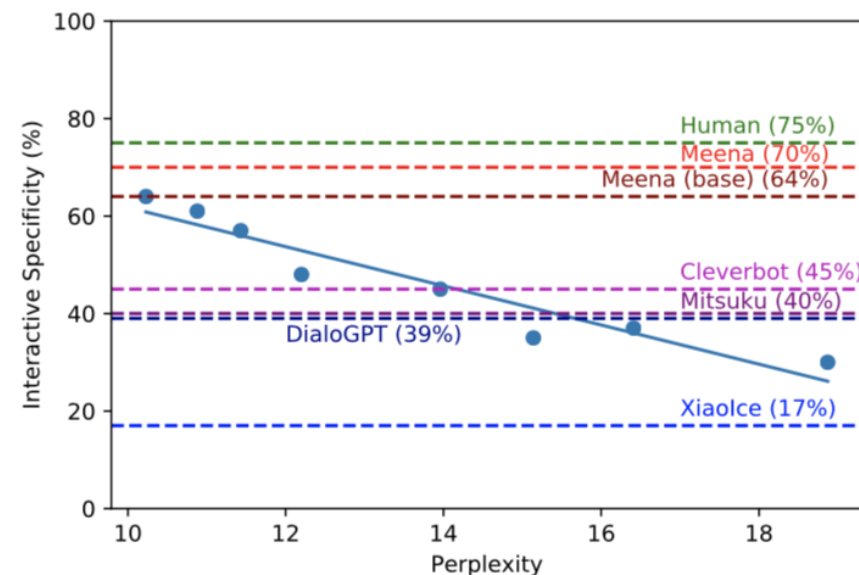


Figure 4: Interactive **specificity** vs perplexity.

## 5. Further Advancing SSA

- 72%  $\pm$  1%, for Meena (base)의 성능을 얻었지만 decoding strategy와 rule 추가로 성능을 79%  $\pm$  1%, for Meena (full)까지 개선해보고자 함

## 5.1 Advancing Decoding

- temperature  $T$ 와 top-k를 다르게 주면서 디코딩 영향을 평가해봄
  - top-k = 40,  $T = 1.0$ ,  $N = 20$ : SSA 72%
  - top-k = 40,  $T = 0.88$ ,  $N = 20$ : SSA 74%
- sample-and-rank에서  $N$ 을 {1,20,400}으로 바꿔가며 평가해봄
  - $N = 1$ 일 때보단  $N = 20$ 일때 유의미하게 좋아짐 (SSA +10%)
  - $N = 400$  일땐 오히려 sensibleness가 안좋아짐

## 5.2 Addressing Cross-turn Repetitions

- Cross-turn Repetitions이란 특정 턴에서 이전 턴의 결과를 반복하는걸 의미함
- 이전결과 반복하기도하고, 답변안에서 모순이 있기도함  
( “I like pizza, but I don’t like it” )
- perplexities가 안좋은 Meena 버전에서 잘 발견되는 현상임 (base 모델에서는 잘 안보이는 현상이긴 함)
- 이를 해결하기 위해 rule을 도입했고 대화 중 2개의 턴에서 long comon sub-sequences를 갖고 있으면 해당 candidates를 제거하게 함
- 이를 통해 SSA 성능이 74% -> 79%로 올라감

Cross-turn Repetition Example 1	
<b>Human:</b>	<i>Hi!</i>
<b>Meena:</b>	<i>Hi! How are you?! :)</i>
<b>Human:</b>	<i>doing good, what would be the best city to visit in Europe?</i>
<b>Meena:</b>	<i>Paris, Barcelona, Amsterdam, Prague</i>
<b>Human:</b>	<i>great list! why?</i>
<b>Meena:</b>	<i>Paris is a beautiful city, and Barcelona is too. And I’ve always wanted to go to Amsterdam and Prague and have never been.</i>
<b>Human:</b>	<i>have you even been to India?</i>
<b>Meena:</b>	<i>No, but I’d love to go.</i>
<b>Human:</b>	<i>how about Japan?</i>
<b>Meena:</b>	<i>I’d love to go to Japan too.</i>
<b>Human:</b>	<i>how about Paris?</i>
<b>Meena:</b>	<i>I’d love to go to Paris too.</i>
<b>Human:</b>	<i>are you lying to me?</i>
<b>Meena:</b>	<i>I’d love to go to Japan too.</i>
<b>Human:</b>	<i>i think you’re lying to me, that makes me sad</i>
<b>Meena:</b>	<i>I’d love to go to Japan, too.</i>
<b>Human:</b>	<i>you’re blocking me out! you have to talk to me</i>
<b>Meena:</b>	<i>I’d love to go to Japan.</i>
<b>Human:</b>	<i>where else?</i>
<b>Meena:</b>	<i>I’d love to go to Japan, too!</i>

Table 5: Multiple near identical repetitions of “I’d love to go to Japan too.”.

## 5.3 Safty Layer

- full Meena 버전의 경우, 내용적으로 민감한 내용들은 filtering mechanism을 적용한 classifier layer를 추가해서 걸렀음 (evaluation 및 실제 대화할 때)

## 6. Related Work

- Human evaluation과 correlation을 갖는 automatic metric을 찾는건 open-domain conversational modeling에서 매우 중요한 goal이었음
- BLEU, ROUGE, 기타 등등이 있었지만 dialog에는 맞지 않음이 밝혀졌었음
- learnable metric을 구축하려는 시도도 있었지만 human labels이 필요하거나 unsupervised approaches를 사용했고 이는 더 복잡하거나 따로 training이 필요했음 (e.g. ranking system)
- 본 연구에서는 any neural seq2seq model에서 사용 가능한 perplexity가 human evaluation과 강한 correlation을 갖는다는 걸 확인함
- DialoGPT 등에서 했던 evaluation setting은 single-turn dialog였지만 본 연구에서는 3 turns까지 가능한 static MTB benchmark와 14 turn까지 가능한 interactive setup으로 평가함

## 7. Discussion

- public domain social media conversations에 대한 perplexity가 good automatic proxy for human judgement (sensibleness and specificity 관점에서)가 될 수 있다는 걸 제안함
- MTB (one to three-turn context)는 first turn에 biased 될 수 있고 context도 많지 않음
  - 주로 다루는 주제도 다음과 같음
  - common sense, basic knowledge, asking/sharing about personality, likes/dislikes, opinions, feelings, hobbies, pleasantries, etc
  - deeper question answering은 불가능 (e.g., how fast is a cheetah)
- Human-likeness란 incredibly broad and abstract concept임
- 이러한 이유로 interactive evaluation을 함 (14 to 28 turns)
- Sensible and specificity 외에 human-like conversation attribute를 확장할 필요가 있음
  - humor
  - empathy
  - deep reasoning
  - question Answering
  - knowledge discussion skills
- Sensible and specificity 경우 sub-component로 쪼갤 수 있음
  - logical
  - personality consistency
  - common sense
  - relevance
  - basic factual correctness
- Future work는 explore the continued optimization of sensibleness via the optimization of test set perplexity