Towards a Human-like Open-Domain Chatbot

(Google Research, Brain Team, 2020)

집현전 중급반 김선행

Towards a Human-like Open-Domain Chatbot

1. Human-like
   - Sensibleness and Specificity Average

2. Open-Domain Chatbot
   - MEENA(Evolved-Transformer)

Sensibleness

- 대화의 맥락에 따라 일관적이고 논리적이며, 사실에 기반한 대화를 나누는가?
- 평가 요소 : common sense, logical coherence, consistency
- 챗봇 대화 예시

  A : 너 운동 좋아해?

  B : 응, 운동 좋아해

  A : 정말? 그러면 어떤 운동을 좋아해?

    B1 : 나 축구를 제일 좋아해(Sensibleness=1)

    B2 : 나 운동 안 좋아해(Sensiblenss=0)

- 한계 : 의문문에는 "I don`t know", 평서문에는 "ok"라는 대답(Sensibleness=1)

  human-like 지표로 불충분

Specificity

- 일반적이고 단조로운 대답이 아니라 진짜 사람처럼 대답을 하는가?
- 챗봇 대화 예시

  A : 난 축구를 좋아해

  B1 : 그렇구나(Sensibleness=1, Specificity=0)

  B2 : 나도, 나는 리오넬 메시 완전 팬이야(Sensibleness=1, Specificity=1)

SSA

- 챗봇의 human-like를 측정하는 방법 : SSA
- SSA = Sensibleness and Specificity Average

- Sensibleness : concrete and basic human quality
- Specificity : more subjective human quality

Static Evaluation and Interactive Evaluation

- 정적 평가(Static Evaluation)
  - 1~3개의 턴으로 이루어진 1,477개의 대화 데이터셋(Mini-Turing Benchmark; MTB)
    1. single turn context : 315, two-turn : 500, three-turn : 662
    2. MTB에는 personality questin 포함(e.g. “Do you like football?”)
  - 성향에 관한 질문 personality question에 대한 일관성 측정 가능

- 동적 평가(Interactive Evaluation)
  - 정적 평가는 데이터셋에 의해 편향이 생길 수 있음
  - 실험자와 챗봇이 1:1로 “Hi”라는 대화와 함께 대화 시작
  - 도메인이나 주제에 대한 제한은 없음
  - 한 회당 최소 14에서 최대 28의 multi turn으로 구성
  - 각 모델마다 100개의 대화 수집, 이를 바탕으로 Sensibleness와 Specificity 평가

SSA consistency

- Crowd working
  - Crowd worker의 일관성을 측정하기 위해 agreement와 Krippendorff`s alpha 사용
  - 5명이 레이블링 하더라도 꽤 높은 일치율을 보임

Metric	Agreement (%)	Krippendorff's alpha
Sensibleness	76 ± 3	0.42 ± 0.03
Specificity	66 ± 2	0.30 ± 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.

SSA vs human likeness

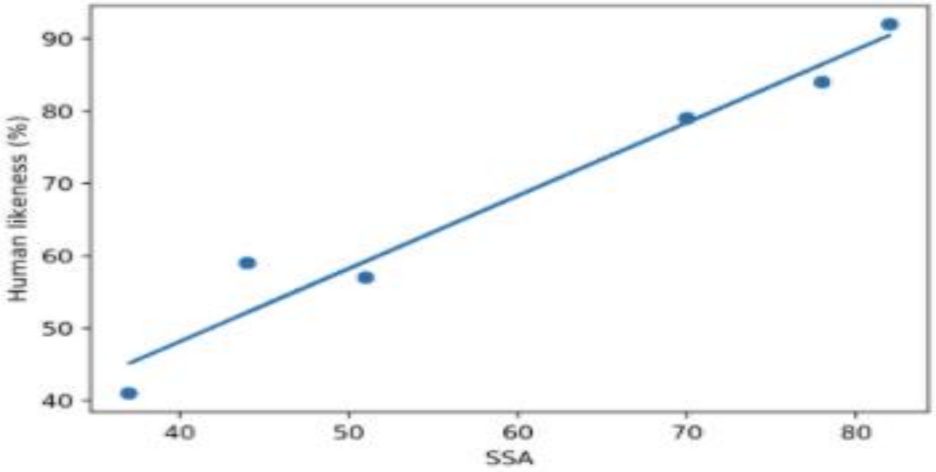


Figure 2: SSA vs human likeness. Each point is a different chatbot, except for the top right one, which is human. A regression line is plotted, for which the coefficient of determination (R^2) is 0.96. The SSA values were collected using static evaluation mode (Section 2.2). The human likeness evaluation 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.

Types of Chatbot

- Close-domain
  - 특정 목표를 달성하기 위하여 keyword 또는 intent에 대한 적절한 반응을 하며 대화가 진행되는 구조

    ex) 여행 관련 예약 시스템, 고객 서비스 시스템

- Open-domain
  - 불특정한 주제의 대화에 대해 사람의 대화와 비슷한 특성으로 대화가 진행되는 구조

    ex) Cleverbot, Mitsuku, Leeluda

Types of Chatbot

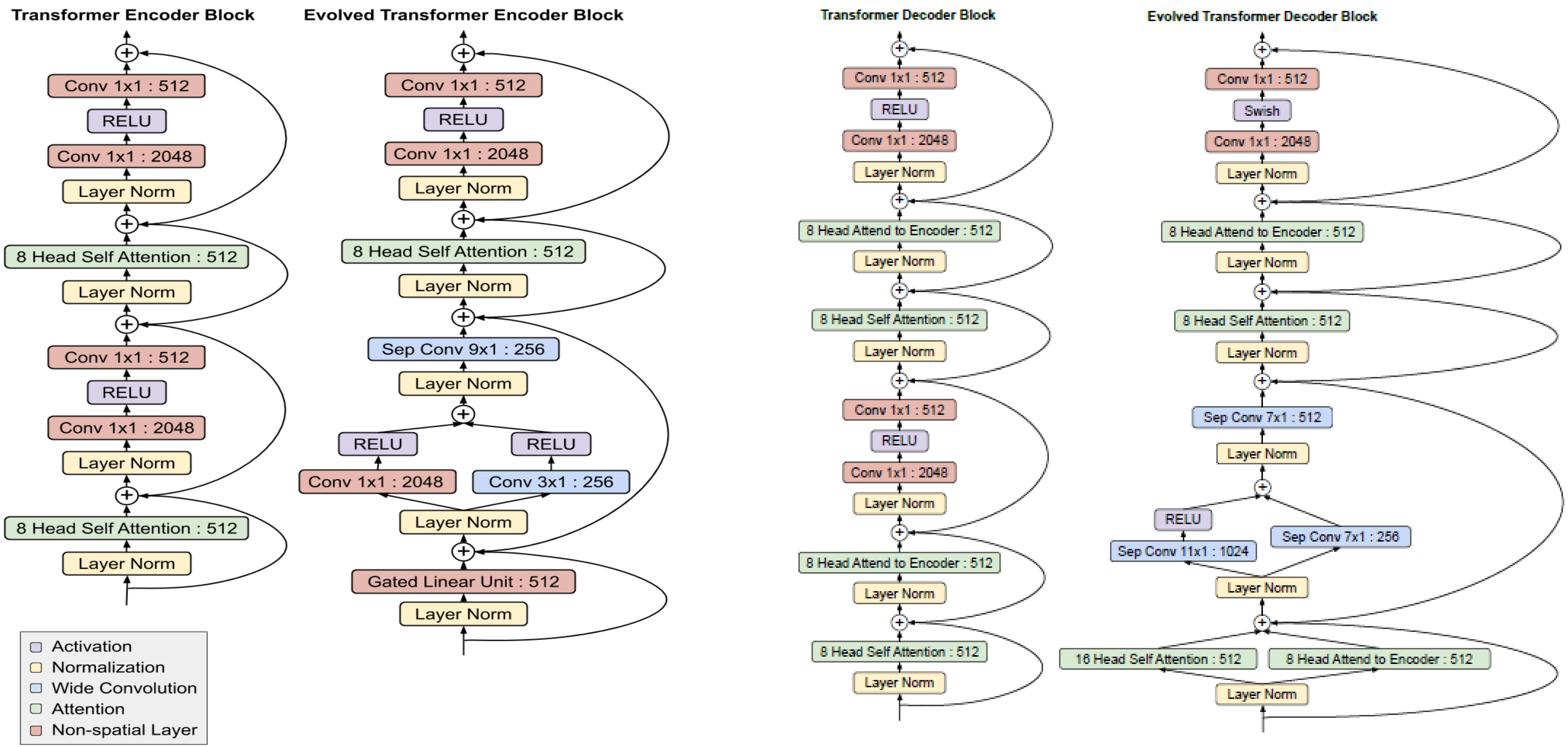
- Complex framework
  - human-designed components
  - language understanding, dialogue management, external communication, response generation

- End-to-End
  - Large neural network models
  - 사람과 같은 대화를 하는 모델을 만드는 방법
    - 학습 데이터와 파라미터를 추가하여 엄청 큰 end-to-end 모델을 만드는 방식
    - end-to-end 모델에 챗봇을 위한 component를 결합하는 방식

Training Data

- Public domain social media conversations을 정제
- source data는 multiple speakers를 포함한 message trees 형태임
- generation quality를 높이기 위해, 다음과 같은 조건을 만족하면 메세지를 삭제함
  1. subword의 개수가 2개보다 적거나 128개보다 많은 경우
  2. 알파벳으로 된 단어들이 70% 미만인 경우
  3. URL을 포함한 메시지의 경우
  4. 발화자의 닉네임이 'bot'을 포함하는 경우
  5. 메시지가 100번 이상 반복되는 경우
  6. 상단 메시지와 n-gram이 겹치는 비율이 높은 경우
  7. 안전하지 않거나 공격적인 메시지가 포함된 경우
  8. 상단 메시지가 인용된 부분은 삭제 또는 삭제된 경우 그에 대한 대답들도 모두 삭제
- 필터링 후에 남은 (context, response) pair의 개수 : 867M
- Vocab size : 8K BPE subwords
- 400억개의 단어 데이터로 이루어진 총 341GB 텍스트

Model Architecture



Training Details

- TPU-v3 pod(2,048 TPU cores)에서 30일간 학습
- Dataset : 40B words(61B BPE tokens)
- TPU setting
  1. TPU-v3의 core는 16GB of high-bandwidth memory를 가짐
  2. 메모리 사용량 최대로하고 각 코어당 8 training examples를 할당
  3. 그 결과 1 step에 1초 정도 소요
  4. full TPU-v3 Pod에서 4M tokens를 1초에 학습시킬 수 있음
  5. 학습이 다 끝났을 때 모델은 164 epoch을 돌았음

Decoding

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. There's a difference.	-3.31

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.

Decoding

- Sample-and-rank
  - Sample N independent candidate response (using plain random sampling with temperature T)
  - candidate response중에서 highest probability를 갖는 걸 final output으로 선택

- Temperature T는 hyper param이고 다음 토큰에 대한 확률분포 p_i를 조절함
- Logits z_i 를 softmax 계산하기 전에 T로 나눔(Hinton et al; 2015)

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

- T값에 따른 변화
  - T = 1, unmodified distribution
  - T가 커지면, contextually rare tokens가 더 많이 출현
  - T가 작아지면, common words가 더 많이 출현하지만 안정적이나 구체적이지 않음(관사, 전치사 등)

Results

	SSA		Sensibleness		Specificity	
	Static	Interactive	Static	Interactive	Static	Interactive
사람	82%	86%	94%	97%	69%	75%
Meena(base)	72%	72%	81%	80%	63%	64%
Meena	78%	79%	86%	87%	70%	70%
XiaoIce	-	31%	-	45%	-	17%
Mitsuku	-	56%	-	72%	-	40%
DialoGPT	51%	48%	62%	57%	39%	39%
Cleverbot	44%	56%	59%	68%	28%	45%

Sample conversation

Conversation B	
Human:	*Hi!*
Meena:	*Hi! How are you this fine evening?*
Human:	*I'm doing well. it's morning my time!*
Meena:	*Good morning! What are your plans for today?*
Human:	*Get my work done then watch a movie at home in the night*
Meena:	*Sounds fun! What movie?*
Human:	*I haven't decided yet, perhaps an Oscar-nominated one. any suggestion?*
Meena:	*I'm not too familiar with recent Oscar winners, but I do know I wanted to see The Grand Budapest Hotel.*
Human:	*ah, tell me about that movie!*
Meena:	*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.*
Human:	*ha! I've searched on the internet and both movies are very colorful!*
Meena:	*I'm looking forward to it. I love colorful movies.*

The Grand Budapest Hotel and Moonrise Kingdom are indeed colorful movies, directed by Wes Anderson.

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

Meena tries to define philosophy.

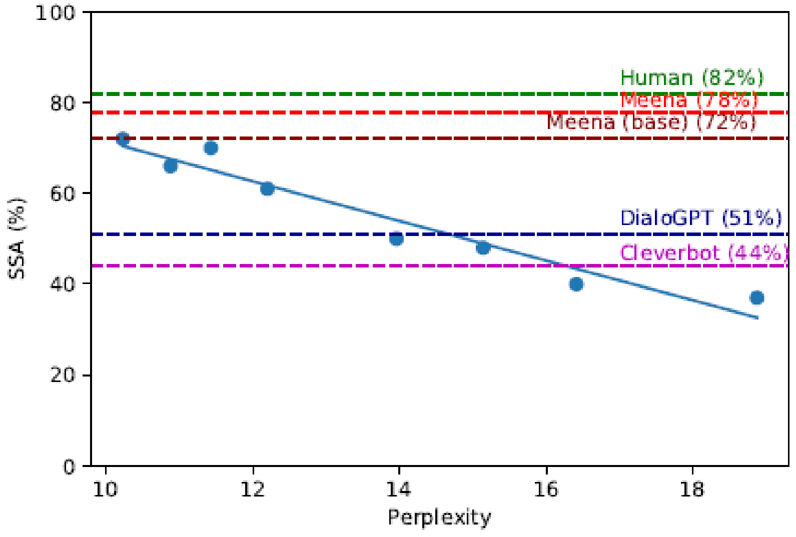
Conversation D	
Human:	*I like how cute cows are*
Meena:	*Ah, interesting. You like how they are cute, right?*
Human:	*yeah, and they're realy smart*
Meena:	*I heard they go to college*
Human:	*cows go to college?*
Meena:	*I heard that a cow went to Harvard.*
Human:	*what did the cow study?*
Meena:	*Bovine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

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

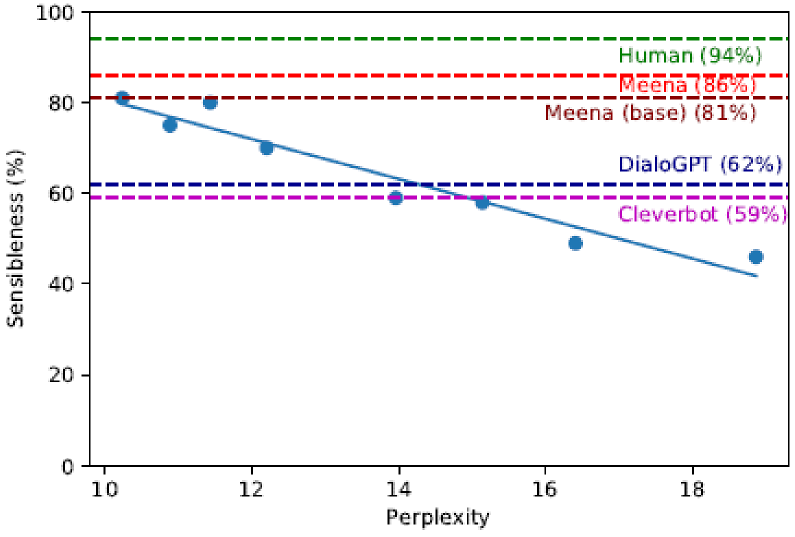
SSA - Perplexity

- Interactive Evaluation

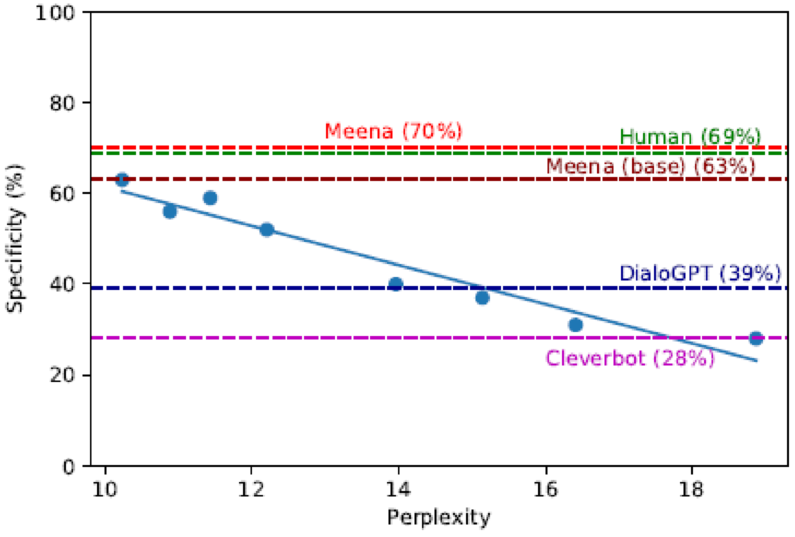
SSA : R^2 = 0.94



sensibleness : R^2 = 0.94



specificity : R^2 = 0.94



Thank you