<https://blog.pingpong.us/meena-presentation/>

**Towards a Human-like Open-Domain Chatbot(2020)**

**Key Takeaways**

End-to-End만으로도 인간에 근접한 기본 대화 능력을 가진 봇을 만들 수 있다.  
즉, 복잡한 프레임워크 없어도 됨(KG 같은 거??), 물론 모델이 충분히 크고, 데이터가 충분히 많다는 전제하에

Sensibleness와 Specificity의 평균인 SSA는 기본적인 대화능력을 측정하는 좋은 평가 메트릭  
2개의 지표지만 인간 대화의 여러 면을 포괄한다. 심지어 레이블링도 쉽다.

또한 많이 쓰는 Perplexity 메트릭이 사람 같은 대화의 Proxy가 될 수 있다.  
PPL이 충분히 낮다면 Sensible & Specific 한 대화 가능  
PPL만 보고 모델 발전시켜도 말을 잘하게 됨(??)

**Closed-domain chatbot:** 키워드와 인텐트가 기반될 수 있음(이 점을 이용해 성능을 높일 수 있다고 생각함), 특정 업무를 수행  
ex) 카카오미니, 샬롯홈

**Open-domain chatbot:** 어떤 토픽에 대해서도 사람같이 이야기를 나눔  
ex) 심심이, Cleverbot ...

**End-to-End vs Complex Frameworks**

End-to-end: 심플하고 단일함, 여러 한계에 노출될 수 있음(Zhang et al 2019)

Complex Frameworks: 기존 대부분의 접근 방식, 여러 세부 모듈이 파이프라인 형태로 붙음

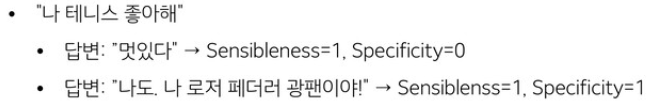
**Sensibleness & Specificity Average(SSA) => 둘 다 0,1 중 하나의 값으로 평가해 평균 냄**

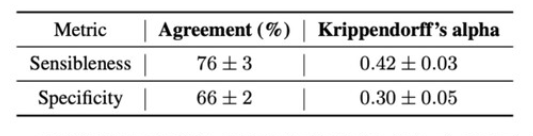
Sensibleness: 대답이 컨텍스트와 함께 말이 되는지 여부, 기본적인 메트릭  
논리적인지 일관적인지 상식적인지 컨텍스트에 맞는지를 따짐  
혼란스럽거나 논리적이지 않거나 맥락에서 벗어나거나 사실이 아니면 0에 가깝게  
인간은 0.97

Specificity  
대답이 구체적인지 평가  
답변이 “잘 모르겠어요”이러면 말은 된다.(Sensibleness 높을 것) 하지만 회피할 뿐이다.(추상적) = 튜링테스트의 한계(재미가 없고 봇인 것을 들키지 않는 전략일 뿐)  
즉, generic ~ 생생한 답변을 구분하며 간단함, crowd sourcing도 쉽다.

**GenericBot vs DialoGPT**

Sensibleness: 70% vs 62%  
Specificity: 0% vs 39%

**Sensibleness(좀 더 확실, 기본적인 척도) -> Specificity(좀 더 주관적인 척도) 순으로 평가함**  


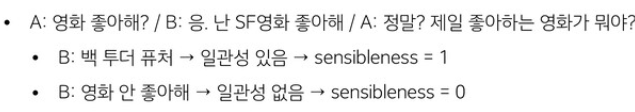


5명에게 레이블(평가)을 받았는데 생각보다 많이 안정적 -> 하지만, 발화들 자체가 주관적이면 평가가 달라지는 경향도 있음

질문이 심플 -> 크라우드 소싱으로 여러 명에게 평가 받을 수 있음

**Static Evaluation**

Mini-Turning Benchmark(MTB): 1477개의 1~3턴의 멀티턴 대화 테스트셋, 컨텍스트(앞의 1~3턴의 대화들, 봇은 그 다음 발화만 말하면 됨)가 정해져있어서 static  
315개의 싱글턴 문장을 Vinyals and Le(2015)와 뢰브너 프라이즈 콘테스트에서 추림  
500개의 2턴과 662개의 3턴 컨텍스트로 만듦  
또한 페르소나와 같이 일관성을 평가하기 위한 데이터도 포함



**Interactive Evaluation**

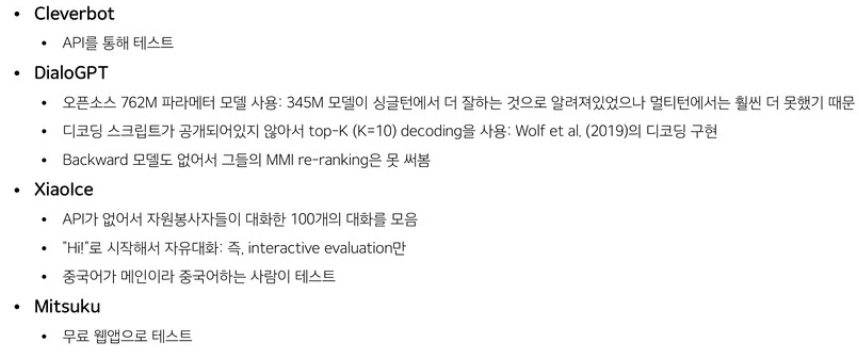
Static은 이미 정해져 있다 = 현실과 bias 존재  
크라우드 소싱(대중들이 그냥 진행)으로 데이터셋 만듬  
Hi -> 자유대화(14~28턴 가짐)  
  
100개의 대화를 통해 평가, 레이블러에게는 챗봇이라고 이미 알려준다.

인간 벤치마크의 평가는 어떻게 하나?

**Human Performance Evaluation**

사람끼리 대화시켜 100개 대화를 평가(레이블러가), 사람끼리 대화하는 걸 이미 알고 시작

다른 모델도 똑같이 진행, 모델을 어떻게 가져왔나?



**높은 수준의 open-domain 멀티턴 모델?** End-to-end 모델에 다른 컴포넌트 없이 데이터와 파라미터 수만으로 학습(end-to-end의 한계(Gao et al. 2019a)가 있음에도 데이터와 파라미터 수로 극복)

**트레이닝 데이터를 어떻게 구했니?** (아마 기존 모델에 파인튜닝할 데이터를 트레이닝 데이터라 칭하는 듯 싶다인건 아닐수도 있다. 아예 생으로 학습할 수도)

1. 소셜 미디어 대화
2. 멀티 파티 트리 형식 대화(단체 카톡방, 소셜 커뮤니티 리플 형태)
3. 최대 7턴까지 대화로 (context, response) 페어를 만듦 => 그런데 학습할 때 (context, response)로 하면은 대화 segment 단위(이 페어는 어떤 대화에 속하는지)를 모르는 것이 아닌가?

**트레이닝 데이터 필터링 조건은 참조**

전체 페어는 8억 6천 7백만개(341GB, 400억개 단어), BPE, sentencepiece lib, 8k vocab  
여러모로 레딧!

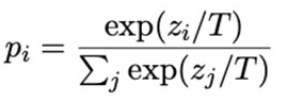
**학습 모델**

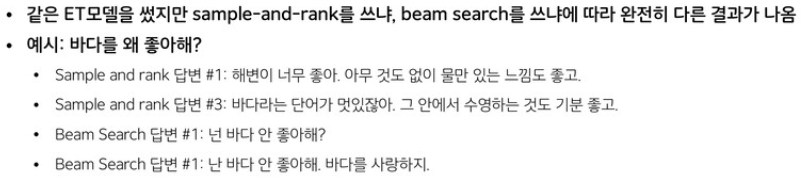
Evolved Transformer seq2seq model(So et al 2019): evolutionary NAS 구조를 트랜스포머에 더한 것, 세부사항 및 학습세팅은 참조

**디코딩 방법**

기존 대화 모델 = 제네릭은 좋지만 구체성은 떨어짐 = 많은 해결책 but 아직 부족  
그렇지만 이번 논문은 PPL이 낮다면(이게 어떤 장점이 있는 걸까 PPL이 낮다면 SSA가 높아지는 것과 연관이 매우 깊나??) Sample-and-Rank decoding 전략만으로 높은 퀄리티 답변 가능

즉, 빔서치가 아니라 sample-and-rank 방식 사용: 토큰을 고를 때 distribution을 따라 랜덤하게 선택, N개의 독립적인 문장 만들고 확률 높은 것으로!

또한, Temperature T에 따라 새로운 토큰을 정할지(T>1), 이전 distribution을 그대로 따를지(T=1), 안전하고 보편적인 토큰을 정할지(T<1)을 선택  




로그 우도(train이든 test든)는 빔서치가 높다.(에러 적음) 하지만, S&R은 재밌는 문장.(PPL이 더 낮은듯?), N=20/T=0.88 사용

**결과**

미나 풀버전 – SSA 79%(사람은 86%) – 기존 챗봇은 가장 높아봤자 56%

또한, 대화가 사람처럼 말하는지에 대한 평가도 SSA와 상관계수가 높았다. 그리고 그 SSA가 PPL과 상관계수가 높았다.(관련도 증가) => PPL을 낮춰 SSA를 높이자

**사람의 SSA Performance**

SSA: static 82, interactive 86  
Sensibleness: static 94, interactive 97, 사람도 잘못 이해하거나, 유머 실패, 컨텍스트 배경 지식 없어서 틀릴 수 있다(비 논리적, 비 일관성)

Specificity: 69/ 75, 아이디어 없거나 흥미 없거나 지식이 없으면 구체적이지 못함

Mitsuku, Cleverbot > DialoGPT > Xiaolce

**SSA 더 발전시키기**  
temperature T, top-K, N 조절(핵심 하이퍼 파라미터 조절)  
interactive 평가시시 1/3의 대화에서 앞에 했던 말 또함, 일관성 문제의 모순도 있음 = 미나는 적게 나타나는 편, 이전 발화에서 특정 턴의 sub-sequence가 높다면 후보에서 제거하여 생성  
또한, 민감한 주제와 위험한 말을 제거하는 classifier가 있다.

**Automatic Metric**

PPL이 결국 짱이다. 다른 시도 있어도..(이전까지 계속 연구되었고 앞으로도 연구될 예정)  
여러 평가 방법도 있다. SSA는 좋지만 empathy같은 부분은 조금 아쉽다.

**테스트셋의 한계**

지협적 질문은 어렵다(치타는 얼마나 빨라, closed-domain이면 가능하지 않을까?)  
1+1은 뭐지? 이것도 빡셈  
따라서, 소셜 대화(튜링테스트 스타일): 일반적인 느낌의 대화일 뿐 지협적 없다. 약간 소개팅 느낌, interactive 테스트가 표본이 짧을 수도 있다.  
따라서, 더 다양한 분야, 다양한 느낌의 테스트셋이 필요하며 그럼 더 객관적으로 모델 평가 가능