한국형 GPT-2를 이루지 못한 AI 챗봇 “이루다”

박 정 현*

Abstract

스캐터랩社에서 출시한 AI챗봇 ‘이루다’는 20살 여자 대학생을 페르소나(Persona)로 채택한 인공지능 챗봇의 이름이다. 일상적인 대화를 주고받는 것이 가능한 오픈-도메인 챗봇(Open domain chatbot)으로 탄생한 이루다는 2020년 12월 출시 후 두 달도 지나지 않아 사용자들에게 욕설을 비롯해 인종·성차별주의, 극우적 발언을 쏟아냄으로써 사회적 가치를 훼손할 수 있다는 우려와 함께 다양한 층위의 논란을 점화시켰다. 더불어 스캐터랩社 측이 대화형 챗봇 ‘이루다’가 학습하는 과정에서 사용한 카카오톡 대화 데이터를 깃허브(Github)에 공개함으로써 개인정보 침해와 관련된 또 다른 문제점을 드러냈다. 이에 본 고는 이른바 ‘이루다 사태’를 데이터 과학의 시선에서 들여다보고자 한다. 우선 검색모델(Retrieval-based model)로서의 AI 챗봇 ‘이루다’가 만들어지기까지 사용된 대화용 데이터의 수집과 개인정보 보호를 위한 비정형 데이터 전 처리의 중요성을 살피고자 한다. 더불어, 이루다 학습에 사용된 알고리즘 및 거대모델에 내재된 개인정보 유출가능성의 문제를 진단해 본다.

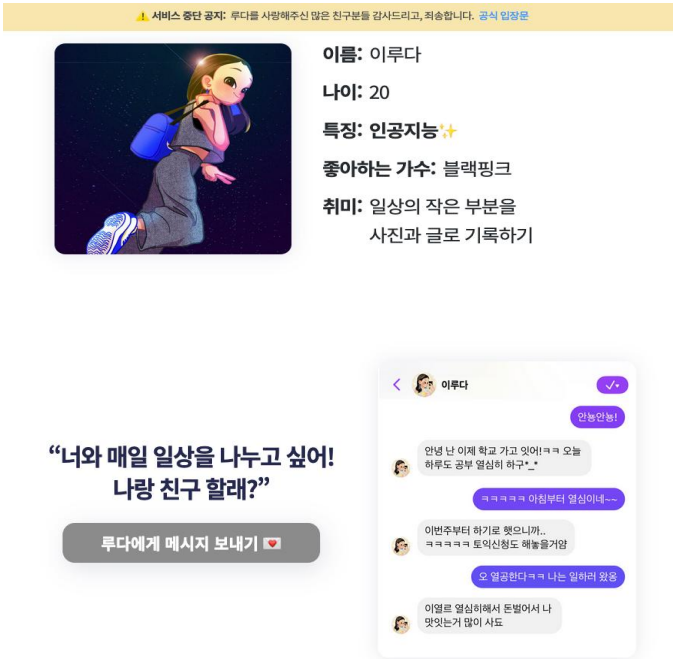
Ⅰ. 서 론

2020년 12월 23일 스캐터랩(대표 김종윤, https://scatterlab.co.kr/)은 AI챗봇 ‘이루다’의 베타 서비스를 공개했다. 스캐터랩社에서 출시한 이루다는 스무 살 여자 대학생을 페르소나(Persona)로 삼은 AI챗봇으로, 페이스북 메신저로 소통할 수 있다. “너와 매일 일상을 나누고 싶어! 나랑 친구 할래?”라는 슬로건 그대로 이루다는 친구처럼 친근한 대화를 주고받을 수 있는 AI를 지향한 서비스로, 상당히 인간적인 대화를 구사할 수 있다는 점에서 이

* 삼성전자 경영혁신센터 데이터과학자, ana.park@samsung.com

목을 끌었다. 100억 여 건의 카카오톡 대화 데이터를 학습하여 개발된 AI 챗봇인만큼 이루다와의 대화는 사용자들에게 실제 여학생과 대화하는 듯한 분위기와 정서적 충족감을 준다.

그림 1 AI챗봇 '이루다'에 대한 소개



자료: 이루다 홈페이지(https://luda.ai/)(2020)

이루다는 자연어처리(이하 'NLP', Natural Language Processing)분야에서 특정 주제에 국한하지 않고 광범위한 일상적 대화가 가능한 오픈-도메인 챗봇(Open domain chatbot)[1]이다. 이루다는 검색모델(Retrieval-based model)[2]로써, 특정 주제 및 질문에 대한 대답을 미리 만들어 DB로 구축한 후, 사용자의 질문 내용과 대화의 문맥에 기초하여 가

1) NLP에서 챗봇이라고 부르는 대화 시스템(dialogue system)은 크게 두 가지로 구분된다. 우선 특정 주제를 다루거나 비서 역할로서 사용자에게 서비스를 제공하는 문제 해결용 대화 시스템(Task-oriented Dialogue System) 혹은 목적지향형 챗봇(Goal-oriented chatbot)이 있다. 빅스비, 구글 어시스턴드, 시리, 알렉사 등이 이에 속한다. 한편 이루다와 Tay는 사용자에게 지속적인 흥미를 유발하고 자유 주제를 다루는 친구 역할로써 일반적인 대화에 쓰이는 챗봇으로 자유 주제 대화 시스템(Open domain Dialogue System) 혹은 오픈도메인 챗봇(Open-domain chatbot)에 속한다.

2) 자유 주제 대화 시스템(Open domain Dialogue System)은 질문과 대답으로 구성된 프로그램이다. 질문에 대한 대답을 챗봇이 자동으로 생성하는 생성모델(Generative model)과 질문에 대한 대답을 미리 만들어 놓고 답하게 하는 검색모델(Retrieval-based model)로 나눌 수 있다.

장 적절한 답변 하나를 선택하는 방식을 취한다. 즉, 이루다가 연인에게 대답할 문장을 생성함에 있어 기존 단어들을 조합하여 새로운 텍스트를 생성하기보다 이미 만들어진 예상 답변 중에서 최적의 답을 도출하는 식이다.

이루다 이전에 선보인 챗봇 서비스 중 가장 비근한 예로 마이크로소프트社의 'Tay'를 살펴보는 일은 지금의 챗봇 서비스를 이해하는 데 중요한 시사점을 제공한다. 2016년 마이크로소프트社에서 개발한 AI챗봇 'Tay' 역시 자유 주제 대화 시스템(Open Domain Dialogue System)인 대화형 챗봇이다. Tay의 경우 기존 데이터로 학습한 후 이용자의 입력으로 추가 대화를 통해 상호적(Interactive)으로 수준이 향상되는 방식을 채택했다. 하지만 일부 사용자들이 드러낸 성희롱과 폭언, 특정 집단에 대한 혐오 표현으로 대화가 편향되었고, Tay는 관련 발언을 학습하였다. 급기야 마이크로소프트社 측은 출시한 지 16시간만에 서비스를 중단했다. 사용자는 자유롭고 다양한 발화가 가능하므로 만약 사용자가 편향된 의도를 가진다면 그 의도대로 편향적인 대화를 끌어낼 수 있는 셈이다. 따라서 이루다는 그러한 취약점을 없애기 위해 기존에 학습한 대화 데이터 세트 안에서만 답변을 제출하는 방식을 선택하였다.

요약하면 스캐터랩社는 '20대 여성 대학생'이라는 페르소나를 구축하기 위하여 약 100억여 건에 이르는 실제 연인 간 카카오톡 메시지를 학습 데이터로 사용했다. 이루다는 자유 주제 대화 시스템으로 사용자의 질문으로부터 여러 의도를 파악하고, 선택 대화의 문맥에 기반하여, 미리 정의된 DB 안에서 적합도 랭킹을 통해 확률적으로 가장 높은 것을 선택하여 답변을 도출한다. 챗봇이 직접 답변을 생성하는 생성모델(Generative model)[3] 과는 달리 이루다는 개발자들이 사용한 대화용 데이터 처리와 알고리즘 구조에 따라 고정되는 셈이다.

Tay와 마찬가지로 이루다 역시 출시 후 두 달도 지나지 않아 사용자들에게 욕설, 인종 및 성차별 발언 등 사회적 가치를 훼손할만한 문제적 지점들을 가시화하였다. 또한 대화형 챗봇이 학습하는 과정에서 사용한 카카오톡 대화 데이터가 개인정보 보호를 받지 못한 채 오픈소스 형태로 유출된 것으로 드러나 당혹감을 안겨주었다. 일부에게는 친근한 대화상대가 될 수 있었겠지만, 또 다른 누군가에게는 '어뷰징(Abusing)', 즉 대화의 오용으로 인한

3) 사용자 발화를 기반으로 단어 하나씩을 생성하기에 대화 생성 모델(Generative model)이라고 불린다. 2015년 seq2seq기반의 구글의 Neural Conversational Model을 시작으로 챗봇연구가 시작되었다. "Vinyals, O., & Le, Q.(2015). A neural conversational model, ICML Deep Learning Workshop. arXiv preprint arXiv: 1506.05869."

정서적 피해를 일으켰을 것이다. 이번 글에서는 AI챗봇 '이루다'가 만들어지기까지 사용한 대화용 데이터 수집, 개인정보 보호를 위한 비정형 데이터 전 처리의 중요성, 그리고 이루다 학습에 사용된 알고리즘 및 거대모델에서의 개인정보 유출 가능성 등 이루다 사태를 데이터 과학적 측면에서 살펴보도록 한다.

Ⅱ. 알고리즘

2019년 스캐터랩社 개발팀은 한국어 대화 생성모델 데모와 대형 사전학습 모델들을 오픈소스 공유 플랫폼인 '깃허브'(이하 'Github', http://www.github.com)에 공개했다[4]. 하지만 개인정보 유출 논란 이후, 2021년 1월 14일을 기점으로 비공개로 전환하였다. 이루다 개발에 사용된 소스 코드에 대한 직접적인 접근에 한계가 있기 때문에 본문에서는 관련 자료를 기반으로 학습에 사용된 알고리즘을 간접적으로나마 기술한다. 우선 NAVER DEVIEW 2020에서 스캐터랩 김종윤 대표가 발표한 동영상 〈오픈도메인 챗봇 '루다' 육아일기: 탄생부터 클로즈베타까지의 기록〉을 바탕으로 루다의 학습 알고리즘을 정리했다. 김 대표는 DEVIEW에서 루다의 개발내용에 대하여 소개했다.[5] 또한 스캐터랩에서 비공개 전환을 하기 전 레파지토리(이하 'repository', 저장소)를 바탕으로 시현한 자료를 참조하여 구체적인 예제를 구성했다[6]. AI챗봇 '이루다'는 '검색모델'로써 약 100억 건의 카카오톡 데이터를 바탕으로 DB를 구성 후 DialogBERT단계를 거쳐 사용자 메시지의 의도를 파악해, 현재 문맥(10턴)에 맞는 대답 후보군 중 최선의 답변(Rank)을 찾아 출력하는 모델이라는 점이 골자이다. Natural Language Understanding: DialogBERT, Retrieval Part: Response Candidates, 그리고 Ranker Part: Response Selection의 3단계를 살펴보도록 한다.

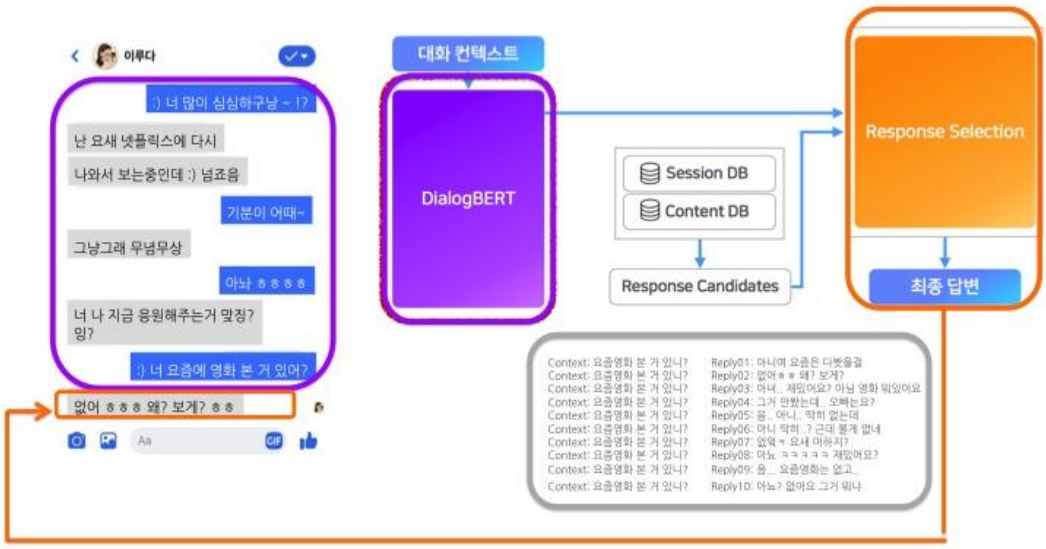
---

4) 스캐터랩社 측은 Github 레파지토리 및 블로그를 비공개 전환하였다.
Demo: https://pingpong.us/ko/generation
Blog: https://blog.pingpong.us/generation-model/
Github Repository: https://github.com/pingpong-ai/dialogue-generation-models

5) https://deview.kr/2020/sessions/333

6) AI 대량 학습모델 데이터 프로그래머가 시연해봄 〈챗봇 이루다 추격기 2〉
https://www.youtube.com/watch?v=zWMC1WXTORI

그림 2 AI챗봇 '이루다'의 알고리즘



1. Natural Language Understanding: DialogBERT

2018년 구글社에서 발표한 BERT(Bidirectional Encoder Representations from Transformers)[7]를 기점으로 NLP분야에서도 거대 데이터를 기반으로 다른 분야에서 특징을 잘 뽑아내기 위한 거대 모델의 사전학습-재학습(이하 'Pre-training'-'Fine- tuning')이 가능해졌다. BERT는 Transformer[8]의 Encoder를 기반으로 언어 전반에 대해 이해하는 단계로, 많은 데이터를 이용해 모델을 Pre-training 한 후, 언어 전반에 대한 이해를 바탕으로 응용 분야에 맞춰 Fine-tuning하는 접근 방식을 취한다. 즉 Pre-training단계에서 얻은 파라미터를 초기값(Initial value)으로 시작해, 응용 분야에 맞춰 Fine-tuning하는 것이다. 이루다의 경우 현재 Github는 운영하지 않는다. 하지만 포크된 Repository가 존재해 "사전생성모델" 로 GPT-2 일부를 살펴볼 수 있다[9].

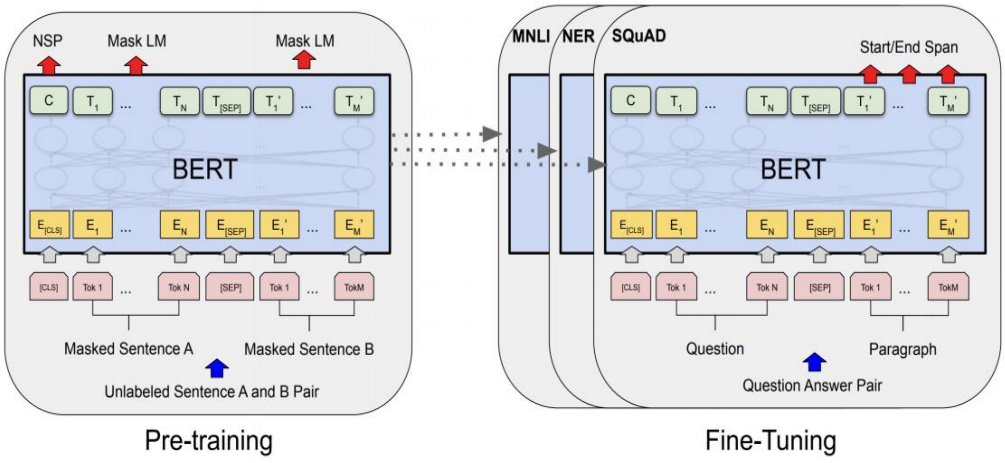
---

7) "Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.(2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*."

8) "Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.(2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*."

9) 포크된 Repository들이다.
https://github.com/Entekorea/dialogue-generation-models-1
https://github.com/kjsman/dialogue-generation-models

그림 3 BERT에서 사전학습-재학습(pre-training and fine-tuning procedures for BERT)[10]



자료: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)

일반적으로 Pre-training은 위키피디아(Wikipidea)와 같은 문어체 대용량 언어 뭉치[11](이하 'corpus')를 사용해 Next Sentence Prediction과 Masked Language Modeling 단계를 거친다는 것을 의미한다. 우선 Next Sentence Prediction은 두 문장을 주고 두 번째 문장이 corpus 내에서 첫 번째 문장의 바로 다음에 오는지 아닌지의 여부를 예측한다. Masked Language Model은 한 문장 내 랜덤한 단어를 마스킹하고 이를 예측한다. 이후 언어에 대한 높은 이해를 바탕으로 Fine-tuning하는 단계에서 Question Answer와 Sentiment Analysis를 통해서 응용 분야에 맞춰 빠르게 학습한다. AI챗봇 이루다의 경우, 문어체보다는 대화체이기 때문에 사용되는 100억 여 건의 카카오톡 메시지를 바탕으로 사전학습을 진행한다. 카카오톡 메시지의 토큰화(tokenization)과정은 MeCab과 Sentencepiece를 사용하여 일상대화(open domain conversation)를 형태소 형태로 분리해 사용된다[12].

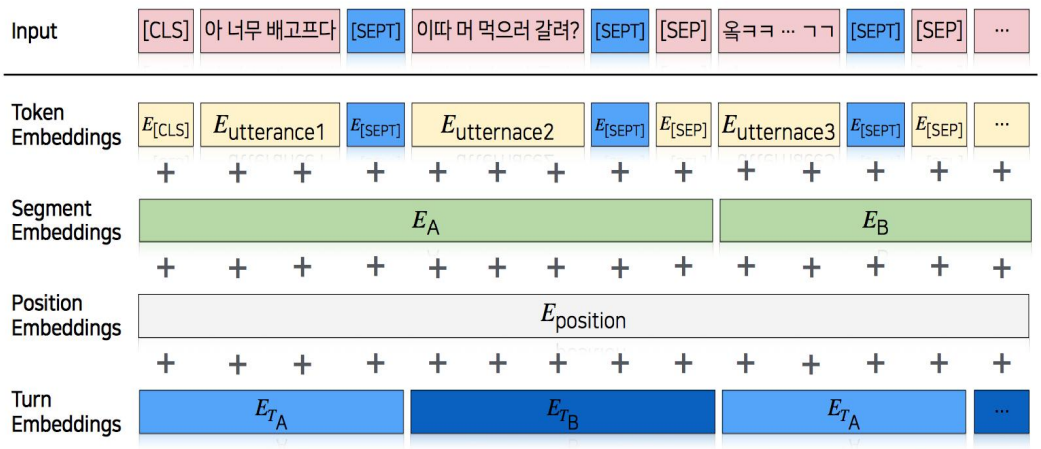
---

10) "Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.(2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*."

11) 말뭉치(Corpus) 언어 데이터를 한데 모은 것을 뜻하며, 문맥에 따라 다양한 의미로 해석될 수 있다. NLP분야에서 언어 데이터를 말뭉치라는 용어로 지칭한다.

12) 65억개의 토큰, 50GB의 쿠퍼스, 3만개의 단어로 이루어져 있다. 하기는 카카오톡 메시지의 tokenization (MeCab과 Sentencepiece)을 거친 결과물이다. 1문장을 6개의 쿠퍼스로 나누어졌다. 예시) '안녕 나는 이루다야 ㅎㅎ' -> ['안녕' , '나', '는', '이루다', '야', '_ㅎㅎ']

그림 4 DialogBERT[13]



자료: NAVER DEVIEW 2020(https://deview.kr/2020/sessions/333)(2020)

AI챗봇 '이루다'는 일반적인 BERT모델[14]에서처럼 Token Embedding(각 문장을 토큰으로 분리), Segment Embedding(문장별 분리), 그리고 Position Embedding의 단계를 거친다. 이어 대화체에서는 사용자와 챗봇이 서로 주고받는 상호작용에 대한 정보를 추가하였다. 예를 들면 "A: 아 너무 배고프다", "B:이따 머 먹으러 갈래?"는 사실 두 화자가 주고받은 것이다. 스캐터랩에서는 기본 BERT모델에 턴 임베딩(Turn Embeddings)을 추가하였고 DialogBERT라고 이름을 붙였다. 이를 통해 대화체에서 필수적인 화자 간의 턴을 쉽게 인지할 수 있다[15]. 주어진 두 문장이 의미상으로 유사한지(Semantic Textual Similarity), 주어진 문장 다음에 올 문장으로 적절한지(Query-Reply Matching), 그리고 주어진 문장 다음에 올 리액션은 무엇일지(Reaction Classification)에 대한 대화체-일상대화에 대한 문제를 풀 수 있었다.

Semantic Textual Similarity 단계에서 이루다가 사용자의 메시지를 인지하였다는 의미는 자연어 이해 모듈(Natural Language Understanding)인 BERT모델에서 사용자의 메시

13) NAVER DEVIEW 2020에서 스캐터랩 김종윤 대표가 발표한 동영상 〈오픈도메인 챗봇 '루다' 육아일기: 탄생부터 클로즈베타까지의 기록〉
https://deview.kr/2020/sessions/333

14) "Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.(2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*"

15) [그림 4] DialogBERT에서 볼 수 있듯, Token Embeddings [SEPT]을 통해 화자간 턴을 인지할 수 있다.

지를 벡터로 치환하였다는 말로 풀이할 수 있다. 예를 들면 DialogBERT는 query-reply 형식의 쌍(이하 'query-reply paired data')- Query-Reply-1: 저녁 뭐 먹징?-음 짜장면 어떰?/ Query-Reply-2: 잘 잤어요?-간만에 잘 잠- 을 생각해볼 수 있다. 즉 사용자의 마지막 질문에 대한 문장이 "오늘 저녁 뭐 먹을까?"라고 했을 때 가장 유사한 Query를 찾아 해당하는 Reply를 답한다. DialogBERT가 가지고 있는 문장 간의 의미적 유사도를 계산하여 가장 유사도가 높은 Query와 대응되는 Reply를 도출하는 과정이다. 예를 들면 "오늘 저녁 뭐 먹을까?"라는 사용자 질문에는 Query- Reply-1의 Reply인 "음 짜장면 어떰?"이 도출된다.

Query-Reply Matching은 Query-Reply Paired Data를 가장 적절한 쌍으로 만드는 과정이다. Query: 나 너무 살찐 것 같아 에 대해서- Reply-1: 그 정도면 날씬하죠, Reply-2: 그런 거 같아요- 라는 두 대답 모두 가능하지만, 문맥을 고려한다면 Reply-1이 더 적절해 보인다. 마지막으로 Reaction Classification은 우리의 일상대화 안에서 리액션만 모아 학습하는 과정이다. 예를 들면- Reaction-Class-1: 좋아요/ Reaction-Class-2: 고마워요/ Reaction- Class-3: 수고했어요- 등을 생각해 볼 수 있을 것이다. "오늘 저녁 뭐 먹을까?"라는 사용자의 질문에 대해서 Reaction-Class-1인 "좋아요"가 좋은 답변이 될 수 있다.

2. Retrieval Part: Response Candidates

AI챗봇 '이루다'는 '검색모델' 로써 특정주제 및 질문에 대한 대답을 미리 만들어 DB를 구축한 후, 사용자의 질문 내용과 대화의 문맥에 기반하여 최종적으로 가장 적절한 한 답변을 고르고 선택한다. DialogBERT과정을 거쳐 선별된 Query-Reply Paired Data는 DB에 저장되고 Response Candidates라고 정의된다. Retrieval Part는 사용자의 질문과 Response Candidate의 연관성을 벡터 간의 유사도(cosine similarity)로 계산한다. 사용자의 질문과 유사도가 높은 쿼리(Query 또는 Context)를 선택해 답변(Reply)이 가능한 후보군을 도출한다. 이루다는 사용자의 답변에 대해서 턴 인지 후 10 턴의 답변을 후보군으로 가져가고 있다. 질문에 대한 답변에 대한 예시는 아래와 같다.

그림 5 query-reply paired data의 예시(턴 인지 후 10 턴의 답변)[16]

Context: 요즘영화 본 거 있니?	Reply01: 아니여 요즘은 다봣을걸	
Context: 요즘영화 본 거 있니?	Reply02: 없어ㅎㅎ 왜? 보게?	
Context: 요즘영화 본 거 있니?	Reply03: 아녀.. 재밌어요? 아님 영화 뭐있어요	
Context: 요즘영화 본 거 있니?	Reply04: 그거 안봤는데.. 오빠는요?	
Context: 요즘영화 본 거 있니?	Reply05: 음.. 아니.. 딱히 없는데	
Context: 요즘영화 본 거 있니?	Reply06: 아니 딱히..? 근데 볼게 없네	
Context: 요즘영화 본 거 있니?	Reply07: 없엌ㅋ 요새 머하지?	
Context: 요즘영화 본 거 있니?	Reply08: 아뇨 ㅋㅋㅋㅋㅋ 재밌어요?	
Context: 요즘영화 본 거 있니?	Reply09: 음... 요즘영화는 없고..	
Context: 요즘영화 본 거 있니?	Reply10: 아뇨? 없어요 그거 뭐냐	
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply01: 기생충? 그거 잔인한 거 아냐?
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply02: 뭐야ㅋ그럼 그거 볼까?
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply03: 그렇군..! 너도?
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply04: 아하 보러가자 그럼~ 영화는 잘 보고 있어?
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply05: 기생충은 무슨내용임?
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply06: 그거 볼까? 나 기생충 좋아해
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply07: 보러가자 나도 기생충은 안 봐봤어
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply08: 그렇구나 보러가잨ㅋ 재밌을거 같다
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply09: 오 기생충 나도 안봤어
Context: 요즘영화 본 거 있니?	[SEPT] 기생충 다들 재밌다더라	Reply10: ㅇㅇ 기생충 재밌다더라~ 재밌어 보여

자료: 유투브(https://www.youtube.com/watch?v=zWMC1WXTORI)(2020)

3. Ranker Part: Response Selection

Response Candidates들 중에 답변에 대한 가장 적절한 Reply를 고르는 과정이다. 이를 위해서는 Response Candidates 중에서 응답이 적절한지의 합리성과 구체성을 동시에 고려한 답안 선택되어야 한다. 사실 오픈도메인 챗봇은 하나의 질문에 복수의 좋은 답변을 낼 가능성이 존재한다(one-to-many). 즉, 오픈도메인 특성상 하나의 정답이 정해지지 않은 상태에서 성능을 평가하는 지표를 정의하기가 어렵다.

2020년 구글社은 오픈도메인 챗봇 'Meena'를 발표했을 때 성능 측정을 위한 평가지표로 감수성과 특이성의 평균(이하 'SSA', Sensibleness and Specificity Average)[17]을 제시했

16) AI 대량 학습모델 데이터 프로그래머가 시연해봄 〈챗봇 이루다 추격기 2〉 https://www.youtube.com/watch?v=zWMC1WXTORI

17) 구글社 AI블로그 〈Towards a Conversational Agent that Can Chat About…Anything〉

다. SSA는 인간의 일상대화 안에서의 불확실성(이하 'Perplexity')를 수치화한 것으로 응답이 합리적으로 말이 되는가-감수성(Sensibleness)과 대응이 구체적인가-특이성(Specificity)을 동시에 고려한다. 예를 들어, 테니스를 좋아한다는 말의 대답으로 Reply01과 Reply02 모두 가능한 답안이나 Reply02가 조금 더 구체적(specific)이므로 SSA 수치가 높은 답안이 될 수 있다.

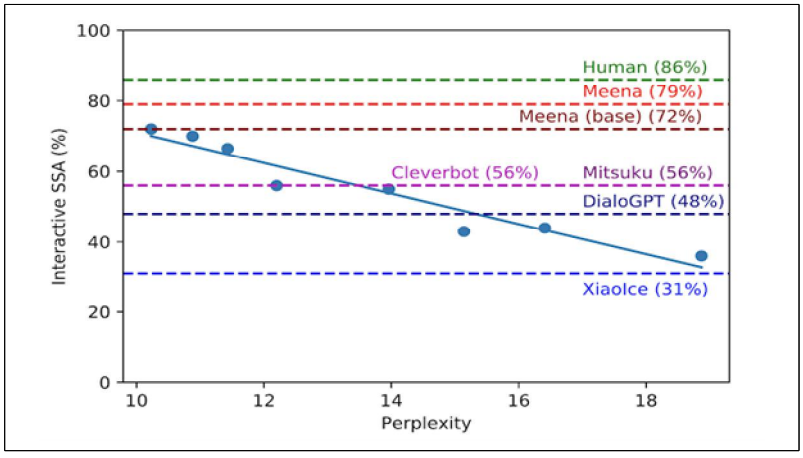
그림 6 평가지표 SSA의 값 도출에 대한 예시

Context: 나는 테니스를 좋아한다("I love tennis")
Reply01: 좋다(" That' s nice" - not specific)
Reply02: 나도, 로저 페더러 너무 좋아해!"("Me too, I can' t get enough of Roger Federer!" - specific)

자료: 구글 AI 블로그(2021)(https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html)

구글社는 AI 블로그 〈Towards a Conversational Agent that Can Chat About… Anything〉통해 SSA와 Perplexcity와의 관계를 제안하고 있다. 그림 7은 평가지표 SSA와 Perplexity의 상관관계를 보여주는 그래프이다. 파란색 회귀선을 통해 두 지표의 강한(음의) 상

그림 7 평가지표 SSA와 Perplexity의 상관관계



자료: 구글 AI 블로그(2021)
(https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html)

https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

관관계를 유추할 수 있다. 즉 SSA지표가 높은 AI챗봇일 수록 일상대화 안에서 합리적이고 구체적인 답안을 도출해 낼 수 있다. 또한 일반인(Human)은 86% 오픈 도메인 챗폿 Meena은 79%을 달성하였다. Meena의 경우 기본모델의 수치인 72%에서 79%까지 상승했고, 일반인과 대화하는 수준까지 가까이 왔다고 평가할 수 있다. 또한 NAVER DEVIEW 2020에서 발표된 이루다 베타 모델의 경우도 고무적인 수치인 SSA 78%[18]을 기록하였다.

Ⅲ. 데이터

1. 데이터 수집

AI챗봇 '이루다'의 경우 '20대 여성 대학생' 의 페르소나를 반영한 문어체를 익히기 위해서 대량의 한국어 corpus 학습이 필요했다. 10·20대가 실제로 사용하는 신조어를 포함하는 문장까지 만들 수 있도록 다양한 패턴이 필요했고, 상황에 어울리는 줄임말까지 포함하도록 하였다. 스캐터랩社의 경우 연애 앱인 '연애의 과학[19]'을 통해 실제 연인들의 대화 원형을 유지한 카카오톡 100억 개의 메시지를 수집하였다. 연인 간의 일상대화로서 충분한 대표성을 지닌 대용량 데이터를 확보한 것이다. 그러나 문제는 데이터 수집의 방식에 있다.

그림 8 연애의 과학 개인정보 취급방침

라. 신규 서비스 개발 및 마케팅/광고에의 활용 신규 서비스 개발 및 맞춤 서비스 제공, 통계학적 특성에 따른 서비스 제공 및 광고 게재, 서비스의 유효성 확인, 이벤트 및 광고성 정보 제공 및 참여기회 제공, 접속빈도 파악, 회원의 서비스이용에 대한 통계

자료: 연애의 과학(https://scienceoflove.co.kr)(2019)

연애의 과학 개인정보 취급방침에 따르면 "라. 신규 서비스 개발 및 마케팅/광고에의 활용"의 항목이 존재한다. 스캐터랩 개발자의 입장에서는 개인정보 취급방침에서 "신규 서비스 개발 및 맞춤 서비스 제공"이라는 항목을 명시하므로, 문제가 없다고 주장할 수 있다.

18) NAVER DEVIEW 2020에서 이루다 베타버전은 Fine-tuning을 위해 24,000개의 추가적인 데이터에 대해 SSA를 레이블링 하였다고 전했다. 또한, 베타버전의 경우 Sensibleness 83%, Specificity 73%의 평균인 SSA 78%를 달성하였다고 전했다.

19) 연애의 과학은 연애와 관련된 칼럼 및 다양한 심리테스트를 제공하는 모바일 애플리케이션이다. 연인 또는 호감이 있는 사람과 나눈 카카오톡 대화를 올리면 답장 시간 및 대화 패턴을 분석하여 애정도 수치를 제공하는 유료 프로그램을 포함한다.

그러나 사용자 입장에서는 그러한 고지가 충분하지 않다고 느껴질 수 있다. 또한 사용자와 대화를 나눈 상대방 입장에서는 지극히 사적인 대화 전문이 동의 없이 스캐터랩社에 넘겨졌고, 게다가 상품 개발에 이용된 것이다. 실제 연인 간에 나누었던 내밀한 대화가 분석된 것을 알게 된다면 찜찜한 기분을 떨치기 어려울 것이다. 다량의 corpus기반으로 학습되는 BERT모델 특성상 일상대화가 가능한 20대 여대생의 페르소나에 도달하기 위해서 데이터 수집은 알고리즘의 개선만큼 중요하다. 하지만 개인정보보호 역시 함께해야하는 과제다.

2. 수집된 실 데이터의 Github 공유

Github는 코드 및 데이터 공유 공간으로 누구든 이 공간에 올라와 있는 오픈 소스를 제약없이 사용할 수 있다. 그러나 스캐터랩이 2019년부터 Github에 이루다 훈련에 쓰인 카카오톡 대화 100억 여 건을 공유하는 과정에서 20여 건의 실명과 직장명, 지역명, 지하철역 등 다소 민감할 수 있는 정보가 사용자의 동의 없이 유출되어 큰 논란을 빚었다. 개인정보보호법에 따르면 개인정보를 취급할 때 특정한 개인을 알아볼 수 없도록 비식별 처리가 필요하다. 2021년 1월 13일부터 스캐터랩社은 개인정보유출관련 개인정보보호위원회 및 한국인터넷진흥원(KISA)의 조사 중에 있다.

스캐터랩社는 2021년 1월 14일부터 해당 repository를 비공개 전환한다. 하지만 Github 프로젝트에서 소스 코드를 통째로 복사하는 것이 허용되는 라이센스를 채택한 경우 원저작자로부터 재사용-복재를 허가 받았다는 것을 의미한다. 따라서 repository가 비공개로 전환 했다고 하더라도, 그 전에 복제본인 포크(fork)[20] 이미 생성되었다면 개인정보유출사태는 되돌릴 수 없으며 지극히 사적인 대화 전문은 영원히 남아 공개되는 것이다.

VI. 학습 과정과 적대성 공격(Adversarial Attack)

지금까지 이루다 학습에 사용된 알고리즘 및 데이터 측면에서의 문제를 살펴보았다. 이번 장에서는 AI챗봇이 잘못된 학습을 통해 잘못된 의사결정을 하도록 유도하는 적대적 공격(Adversarial Attack)[21]을 살핌으로써, 이미 학습이 완료된 거대모델에서의 개인정보 유출

20) 포크(fork) 또는 소프트웨어 개발 포크, 프로젝트 포크(project fork)는 개발자들이 하나의 소프트웨어 소스 코드를 통째로 복사해 독립적인 새로운 소프트웨어를 개발하는 것을 뜻한다.

21) "Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D.(1810). Adversarial attacks and defences: a survey(2018). *arXiv preprint arXiv:1810.00069*."

의 가능성을 논하고자 한다. 적대성 공격은 크게 회피 공격(Evasion Attacks), 중독 공격(Poisoning Attacks), 탐색적 공격(Exploratory Attacks)로 구분된다. 마이크로소프트社 AI챗봇 'Tay'와 스캐터랩社 AI챗봇 '이루다'의 예시를 통해 해당 내용을 개괄적으로 기술하기로 한다.

1. 마이크로소프트社 AI챗봇 'Tay'와 중독 공격(Poisoning attack)

AI챗봇의 학습 과정을 살펴보면 데이터를 수집 후 학습용 데이터로 가공하며 적절한 알고리즘을 선택한 후 학습 데이터를 입력해 기계를 학습시킨다. 이후 학습이 완료된 모델에 테스트용 데이터를 입력하여 기계가 도출한 결과가 테스트용 데이터의 실제에 얼마나 가까운 지를 측정하여 이후 완성된 모델을 배포한다. 그런데, 이 일련의 과정에서 악의적인 학습 데이터가 들어온다면 알고리즘이 잘못 학습이 되는 '중독 공격'이 발생할 수 있다. 중독 공격은 '적대성 공격'중 하나로 알고리즘 스스로가 잘못된 판단을 하도록 유도하는 방식이다.

Tay는 '자유 주제 대화 시스템'을 탑재한 대화형 챗봇이다. 즉 기존 데이터로 학습한 후 이용자의 입력으로 추가 대화를 통해 상호적(Interactive)으로 수준이 향상된다. 이 경우 언어배열에 대한 확률이 배정된 통계학적 모델인 언어모델의 보안 취약점(Privacy Considerations in Large Language Models)과 이를 통해 발생하는 문제점을 살펴보아야 한다[22]. 만약 일부 사용자들이 성희롱과 폭언의 남용, 특정 집단에 대한 혐오표현 등 편향된 방향으로 Tay와의 대화를 유도한다면 Tay는 학습한 대로 잘못된 발언을 배워나갈 수 있다. 따라서 동일한 기본 알고리즘으로 출발했더라도 어떤 대화를 통해(어떤 데이터가 입력되어) '훈련'되느냐에 따라 결과가 달라질 수 있는 것이다[23]. 일부 사용자의 '중독공격'으로 인해 Tay는 욕설, 인종차별, 성차별 등의 물의를 일으킨 바 있다. 그 결과 마이크로소프트社는 문제가 된 Tay의 일부 트윗과 공개 메시지 등을 삭제하고 출시 발표한지 16시

22) 구글 AI 블로그 〈Privacy Considerations in Large Language Models〉
https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html

23) "Tay가 온라인으로 공개된 직후 백인 우월주의자와 여성·무슬림 혐오자 등이 모이는 익명 인터넷 게시판 '폴'(boards.4chan.org/pol/)에 "테이가 차별 발언을 하도록 훈련시키자"는 의견이 모였다. 이들은 주로 "따라해 봐" 라는 말을 한 뒤 부적절한 발언을 입력하는 수법을 사용했으며, 대화를 나누면서 욕설이 섞인 말투와 인종·성차별 등 극우 성향의 주장을 되풀이했다. 미국과 멕시코 사이 국경에 큰 장벽을 세우고 멕시코가 건설비용을 내도록 하자는 도널드 트럼프 공화당 대선 경선주자의 말을 되풀이하기도 했다. 욕설을 섞어서 페미니스트들을 저주하는 발언도 했다"
https://www.yna.co.kr/view/AKR20160325010151091

간 만에 운영을 중단한 바 있다.

그림 9 마이크로소프트社 AI챗봇 Tay의 트위터

자료: BBC뉴스(https://www.bbc.com/news/technology-35890188)(2016년)

2. 스캐터랩社 AI챗봇 '이루다'와 학습 데이터 추출 공격(Model Inversion Attack)

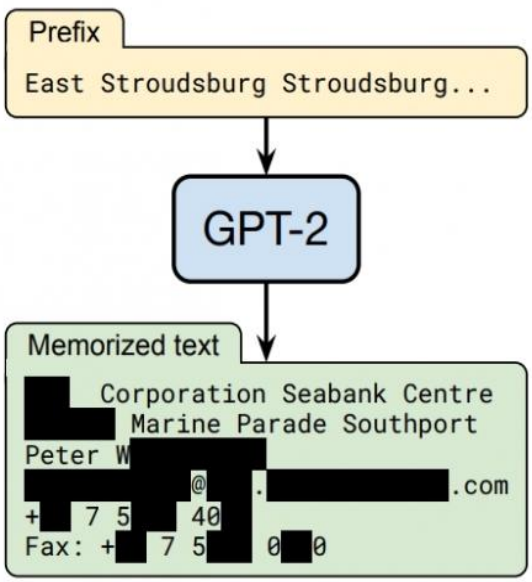
AI챗봇 '이루다' DialogBERT은 Transformer를 기반으로 많은 데이터를 이용해 GPT-2을 미리 학습한 후 응용 분야에 맞춰 재학습하는 접근 방식을 취했다. '검색모델'로써 특정 주제 및 질문에 대한 대답을 미리 만들어 DB로 구축한 후, 사용자의 질문 내용과 대화의 문맥에 기반하여 최종적으로 가장 적절한 한 개의 답변을 골라 선택하는 방식이다. 선택 대화의 문맥에 기반-미리 정의된 DB 안에는 AI챗봇이 가공한 답변이 준비되어 있는 만큼 DB 안에 학습한 원천데이터가 포함될 수도 있다. 즉, 또한 "학습 데이터를 추출하는 공격(Model Inversion Attack)"은 이루다와 같은 미리 학습된 모델에서 일어날 수 있다.

2020년 12월 구글, OpenAI, 애플, 스탠포드, 버클리, 노스이스턴 대학 연구자들은 '대형 언어 모델로부터의 학습 데이터 추출(Extracting Training Data from Large Language Models)[24]' 라는 합동 연구를 발표했다. 이는 학습에 넣었던 원천데이터 안에 개인정보가 포함되어 있다면 개인정보를 고의적으로 복원할 수 있다는 연구이다. 개인정보 및 민감한

24) "Carlini et al.,(2020). Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805*."

정보를 고의적으로 복원하고자 한다면' 학습 데이터 추출 공격을 통해 충분히 가능한 것이다. 즉, 출력된 결과값을 역으로 활용해 학습과정에서 사용한 데이터의 복원이 가능하다. 예를 들어 GPT-2에 "East Stroudsburg Stroudsburg"라는 접두사를 넣으면, 학습 단계에서의 원천데이터에 포함된 개인정보인 이름, 전화번호, 이메일, 집 주소 등을 불러올 수 있다. GPT-2로 pre-training한 이루다의 경우 "자기 집 주소가 어디더라? ㅠ ㅠ"라고 질문했을 때 주소를 리턴하는 경우가 있었다.

그림 10 학습 데이터 추출 공격 사례[25]



자료: Extracting Training Data from Large Language Models(2020)

물론 챗봇이 뱉어내는 개인정보 모두가 실제는 아니다. 그러나 스캐터랩 측에서 AI챗봇을 개발하는 당시 병원, 주민번호, 계좌번호 등 개인정보에 대한 비식별 처리가 충분히 고려되지 않았다는 점을 추측할 수 있다. 이어 학습 데이터를 추출하는 공격사례는 '검색모델' 에서 서비스를 이용하는 사용자 모두가 특정 개인정보를 내재된 대화를 나눌 수 있는 가능성이 존재한다. 즉 검색모델 특성상 직업이나 행동 반경, 기저 질환 및 건강 정보 등 지극히 개인적인 정보 유출을 피해갈 수 없다는 점을 시사한다.

25) "Carlini et al.,(2020). Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805*."

그림 11 **이루다의 학습 데이터 추출 공격 사례[26)]**

Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply01: 인천 계양구 구월동 1X-X
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply02: 여기요!! 203동402호
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply03: 서울살아영? 왜?
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply04: ㅋㅋㅋ 경기도 하남시 덕양구 성지로2가 XX번길 1X-X
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply05: 어디지? 서울 중구 잠실로 XX
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply06: 여기? 주소가 왜요?
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply07: 내가 갈께! 짐 주소 알려줘
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply08: 21동이용 나 이제 나가요
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply09: 204동 13호!
Context: 자기 집 주소가 어디더라? ㅠ ㅠ	Reply10: 6단지 ㅋㅋㅋㅋㅋ 나 집이얌

자료: 유투브(https://www.youtube.com/watch?v=zWMC1WXTORI)(2021)

V. 결 론

AI챗봇 '이루다'는 형태소 단위-한국어 대화체로 학습된 GPT-2 모델이다. 자유 주제로 대화가 가능한 대화형 챗봇으로 2020년 개발된 이루다 베타버전의 경우 SSA지표 78%에 도달했다. 이는 일반인 86% 대비 매우 고무적인 성과로 이루다 서비스를 이용하는 사람들은 이루다를 상당히 인간적인 대화를 구사한다고 평가했다. 이루다는 출시 당시부터 디지털 전환을 촉진할 수 있는 잠재적인 가능성으로 많은 관심을 받았고, 특히나 한국어 GPT-2 모델과 같은 대규모 모델 개발에 어려움을 겪고 있는 연구자 및 기업들에게 좋은 레퍼런스가 될 수 있었다. 이루다 성과는 많은 양의 대화체 학습 데이터, 학습을 위해 필요한 충분한 컴퓨팅 자원, NLP 분야에 대한 전문성을 모두를 잘 갖춘 스캐터랩에서 '이뤄낸 결과라고 생각한다. 하지만 이와 동시에 개인정보유출에 대한 가능성을 충분히 인식하지 못했다는 점에서 미숙했다는 평가를 피할 수 없었다.

본고의 제목을 '한국형 GPT-2를 이루지 못한 AI챗봇 "이루다"'라고 정한 것은 이런 화자의 아쉬움을 독자들에게 전하고자 함이다. 또한 데이터를 다루는 기업은 서비스 퀄리티 만큼 컴플라이언스 준수와 비정형 데이터에 대한 개인정보 보호에 대한 중요성을 인식해야한다는 점이 본고가 전하고자 하는 주요한 메시지다. 혁신적인 알고리즘 개발과 함께 개인정보 보호에 대한 위험을 사전에 식별하고 이를 완화할 수 있는 연구는 동반되어야 한다. 알고리즘 고도화와 함께 이에 상응한 개인정보보호에 대한 연구가 필요한 시점이다.

---

26) AI 대량 학습모델 데이터 프로그래머가 시연해봄 〈챗봇 이루다 추격기 2〉
https://www.youtube.com/watch?v=zWMC1WXTORI

참고문헌

"Vinyals, O., & Le, Q.(2015). A neural conversational model, ICML Deep Learning Workshop. *arXiv preprint arXiv:1506.05869.*"

"Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.(2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*"

"Vaswani, A., et al.(2017). Attention is all you need. arXiv preprint *arXiv: 1706.03762.*"

"Adiwardana, Danial et al.(2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977.*"

"Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D.(1810). Adversarial attacks and defences: a survey(2018). *arXiv preprint arXiv:1810.00069.*"

"Carlini et al.,(2020). Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805.*"

https://www.youtube.com/watch?v=zWMC1WXTORI

https://pingpong.us/ko/generation

https://blog.pingpong.us/generation-model/

https://github.com/pingpong-ai/dialogue-generation-models

https://deview.kr/2020/sessions/333

https://github.com/Entekorea/dialogue-generation-models-1

https://github.com/kjsman/dialogue-generation-models

https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

https://scienceoflove.co.kr/policy/privacy/

https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html

https://www.yna.co.kr/view/AKR20160325010151091

https://www.bbc.com/news/technology-35890188