대형 언어 모델을 활용한 대화 시스템의 최근 동향

NAVER AI Lab ▌신재민

1. 서 론

흔히 말하는 챗봇 (Chatbot) 은 크게 오픈 도메인 대화 시스템 (Open-Domain Dialogue System, **ODD**) 과 목적 지향형 대화 시스템 (Task-Oriented Dialogue System, **TOD**) 로 나뉘어져 있다. 위 그림1에서 보다시피 전자는 이용자와의 대화를 더욱 오래, 재미있게 하는 일상 대화가 목적이고 한국에서 대중적으로 알려진 서비스로는 심심이, 이루다 등이 있다. 후자는 쇼핑이나 날씨 안내, 비행기와 호텔 예약 등 사용자가 원하는 특정한 문제를 해결하는데 특화된 챗봇이고 삼성 Bixby, Google Assistant, Apple Siri, Amazon Alexa 등이 가장 널리 알려진 서비스 예시다.

ODD와 TOD 모두 깊은 역사를 가지고 있는 연구 분야들이지만 본 논문에서는 주로 현재 빠르게 발전하고 있는 GPT-3 [1], Gopher [2], T5 [3]와 같이 대규모의 언어 데이터로 사전 학습된 대형 언어 모델 (Large Language Model, **LLM**) 들의 등장 이후로 어떻게 각 분야의 연구 동향이 이에 발맞추어 바뀌어 가고 있는지 소개한다.

2. 오픈 도메인 대화 시스템 (Open-Domain Dialogue Systems)

오픈 도메인 대화 시스템에 관한 연구는 상당히 오래 전부터 되어왔고 이에 대해서는 [4, 5, 6] 에서 심도 깊게 다루고 있다. 그러나 대규모로 사전 학습된 대형 언어 모델의 빠른 발전에 맞추어 대형 대화 시스템이 도래하면서 ODD 연구도 매우 빠른 속도로 발전하고 있다. 이번 섹션에서는 이와 관련해서 어떠한 발전들이 있었는지 포괄적으로 다루고 있고, 급격하게 좋아진 모델들의 대화 성능으로 인한 부작용들을 해소하기 위한 노력들에 대해서도 간략히 소개한다.

2.1. 대형 대화 시스템의 도래 (The Emergence of Large Pre-trained Dialogue Models)

2019년 상반기에 OpenAI 에서 GPT [7] 의 후속으로서 모델 사이즈가 최소 10배는 더 커지고 10배 이상의 데이터로 학습시킨 GPT-2 [8] 를 발표했다. 그 모델의 구조와 학습 방법론을 그대로 채용해서 2019년 말에 Microsoft 에서 발표한 모델이 최초의 (공개된[1]) **대형 사전 학습 대화 시스템**이라고 할 수 있는 DialoGPT [9] 이다. DialoGPT는 약 1.5억 건의 (147M) "단일 턴 (Single-turn)" Reddit 스레드 데이터에 학습시켰고 GPT-2 large (1.5B) 의 절반 정도 크기인 (그리고 T5 Large [3] 와 같은 모델 크기인) DialoGPT large (762M) 모델을 오픈소스로 공개[2]하였다. 이후 DialoGPT 는 다양한 ODD 연구의 핵심적인 Foundation Model [11] 로서 사용되었다. 그 이전에도 HuggingFace 가 NeurIPS 2018 ConvAI2 대회에서 발표했던 TransferTransfo [12] 라는 모델이 있었지만, DialoGPT 보다 훨씬 작은 규모로 BERT-base (110M) 정도의 모델이었다. 최근에는 (2022년 5월) Microsoft 에서 이 연구의 후속작으로서 GODEL [13] 이라는 논문을 발표하고 모델을 공개[3]했는데 이 모델은 약 5.5억 건의 (550M) "멀티 턴 (Multi-turn)" Reddit 스레드 데이터로 학습시켰고 이 외에도 약 5백만 건 (5M) Instruction과 지식 기반 (Knowledge-grounded) 대화 데이터를 추가로 학습시켰다. 대화 데이터의 양과 모델 크기 외에도 이전 버전인 DialoGPT와 특별히 달라진 점을 나열하자면 학습에 사용된 데이터의 종류와 모델 구조일 것이다. 우선 학습에 사용된 데이터는

* 종신회원

1) 이전에도 Microsoft 에서 XiaoIce [10] 라는 다양한 모듈들이 붙어 있는 비교적 큰 대화 모델을 만들어 서비스했지만, 소스 코드와 학습된 모델이 공개된 적은 없었다.

2) Large 모델 외에도 Small (117M) 과 Medium (345M) 도 함께 공개하였다. https://github.com/microsoft/DialoGPT

3) https://github.com/microsoft/godel

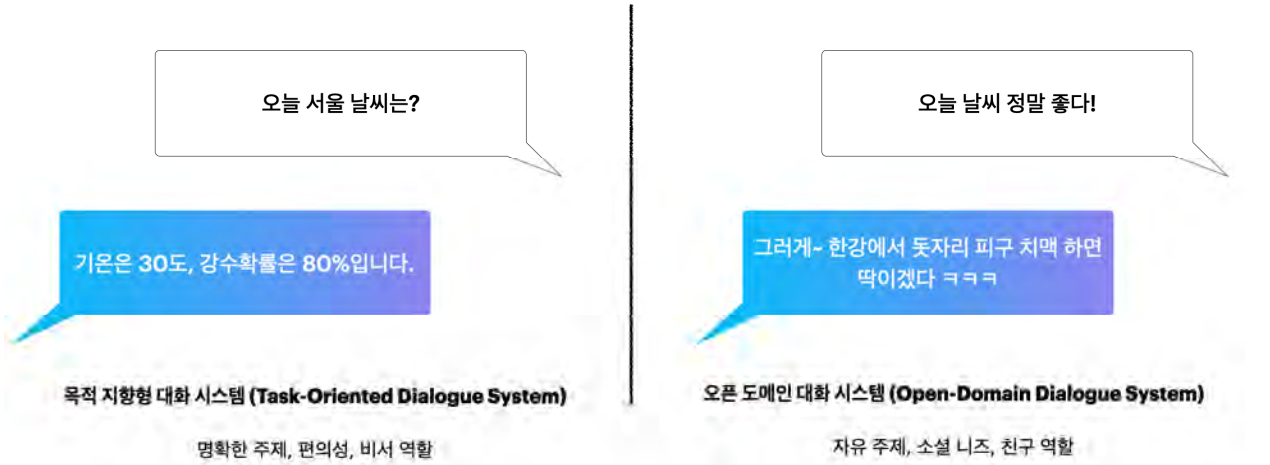

그림 1 목적 지향형 대화 시스템 vs 오픈 도메인 대화 시스템의 예시. 같은 날씨라는 주제를 가지고도 진행 양상이 다른 것을 볼 수 있다.

Reddit 말고도 질의 응답 (Question Answering) 데이터인 MS-MARCO [14] 그리고 UnifiedQA [15], TOD 데이터인 DSTC7-Task-2 [16] 와 Schema-Guided Dialogue [17] 데이터를 학습에 사용했다고 한다. XL 모델의 경우 GPT-J (6B) [18] 를 차용하고 공개했다. 기존의 DialoGPT 모델보다 GODEL이 대화형 질의응답 (Conversational QA), TOD 등 특히 외부 지식을 요하는 다양한 종류의 문제들에서 훨씬 더 높은 성능을 보였고 앞으로 다양한 대화 도메인에서 사용할 수 있는 좋은 Backbone 모델이 될 것으로 기대된다.

Google 에서는 2015년에 학계 최초로 Sequence-to-sequence [19] 모델 구조를 대화 모델로서 활용하는 연구를 발표했었고 2020년 초에 Meena [20] 라는 모델과 논문을 발표한다. 이 논문에서 눈 여겨 볼 점들은 크게 세 가지다. 우선 Evolved Transformer [21] 라는 유전 알고리즘 기반 신경망 아키텍처 탐색을 적용했고, DialoGPT 보다 약 3.5배 정도 큰 모델 (2.6 B) 로 341GB의 데이터에 학습하였다. 그리고, Sensibleness and Specificity Average (SSA) 라는 새로운 수동적 대화 평가 방법론을 (Human Evaluation Protocol) 제시하였다. SSA가 가장 중요하게 여기는 점은 두 가지로 "대화가 문맥상 말이 되는가 (Sensibleness)"와 "대화가 구체적인가 (Specificity)" 를 평가자에게 질문한다. 여기서 주목할 점은 SSA 라는 수동적인 평가 지표가 흔히 쓰이는 대화 모델의 자동화된 평가 지표인 Perplexity (PPL) 와 매우 높은 상관관계를 보여주었다는 것이다. 이 점은 추후 나올 대화 모델들의 스케일을 키우는 중요한 근거가 된다. 이후 2022년 초에 Google 에서 LamDA [22] 라는 논문을 발표하는데 이번에는 모델의 크기를 50배 이상인 137B 까지 증가시켰다. 이에 따라 위의 SSA 지표를 사람에 더욱 가깝게 비약적으로 발전시켰다. 또한, LamDA 저자들이 특히나 핵심적으로 강조한 것은 안전성 (Safety) 과 근거성 (Groundedness) 이다. 이 두 문제를 해결하기 위해서 저자들은 크라우드소싱을 통해 각각에 해당하는 대화 데이터를 수집하였고, 사전 학습된 대화 모델을 수집된 데이터에 학습시켰다. 다 합쳐서 몇 만 건 되지 않는 데이터임에도 불구하고 여러 가지 지표에서 훨씬 높은 성능을 보여줌으로써 그 효용을 증명했다. 이처럼 매우 중요한 문제 의식을 가지고 훌륭한 성과를 보인 두 논문의 가장 아쉬운 점은 다른 기관들에서 발표한 대형 대화 모델들과 달리 사전 학습된 모델과 코드 모두 "비공개" 상태라는 점 때문에 외부 기관에서 재현이 어렵고 관련된 후속 연구를 하기 어려운 상황이다.

Meta 에서도 꾸준하고 활발하게, 그리고 투명하게[4] 대화 모델 연구를 해왔고 그 첫 번째 노력의 결실이 2020년에 Google 의 Meena 발표 직후에 나온 BlenderBot [23] 이다. BlenderBot 의 경우 Meena (2.6 B) 의 3.6 배 정도 되는 9.4 B 모델로 발표 당시에는 가장 큰 대화 모델이었다. 그러나 BlenderBot 이 주목받았던 것은 단순히 사이즈의 문제가 아니라 그동안 Meta (당시 Facebook) 에서 해왔던 다양한 ODD 관련 연구 성과들을 하나로 모았기 때문이다. 우선, Meta 에서는 그동안 파편적으로 공개했던 데이터셋들인 공감 대화 (Empathetic Dialogues) [24], 페르소나 대화 (Persona Chat) [25], 그리고 지식 기반 대화 (Wizard of Wikipedia) [26] 를 하나로 **섞은** Blended Skill Talk [27] 이라는 대화 데이터셋을 공개하였다. 이 데이터

4) ParlAI Projects 페이지에는 여태까지 Meta가 BlenderBot-1,2,3 프로젝트들을 만들기까지 해왔던 모든 연구들이 공개되어 있다. https://parl.ai/projects/

에 fine-tuning 하는 것이 수동적 대화 평가 지표에서 큰 상승 효과를 주었다. 또한, 2019년에 제안했던 ACUTE-Eval [28] 이라는 새로운 수동적 대화 평가 방법론에서 Meena 를 포함한 기존 챗봇들보다 좋은 성능을 보였다. 이후 2021년에는 BlenderBot 의 스케일을 키우는 대신에 두 가지 핵심적인 기술을 더한 BlenderBot 2 [29, 30] 을 발표했다. 우선 기존에 Facebook 에서 발표되었던 Retrieval Augmented Generation [31] 과 효율적인 검색 기술인 Fusion-in-Decoder [32] 을 바탕으로 인터넷을 검색하는 대화 모듈 [29] 을 장착했다. 이는 앞서 언급했던 GODEL 과 LamDA 와 비슷한 동기를 가진 방향의 발전이다. 사람이 직접 비교 평가했을 때 BlenderBot 1 에 비해서 같은 모델 크기임에도 불구하고 훨씬 더 발전된 성능을 보인 것이 높게 평가된다. 마지막으로, 2022년에 BlenderBot 3 [32] 이 발표되었는데 먼저 주목할만한 점은 GPT-3의 스케일로 BlenderBot 2 모델을 175B 으로 확장했다는 점이고 이를 위해 최근에 공개한 LLM인 OPT-175B [33] 를 Backbone 모델로 사용했다. 전작에 비해 BlenderBot 3 에서 저자들이 특히 강조한 부분은 Continual Learning [35,36] 과 LamDA 와 마찬가지로 안정성 (Safety) [36,37,38] 이다. 이들이 여기서 말하는 Continual Learning 은 일반적인 평생 학습 알고리즘을 말하는 것이 아니고 배포된 데모에서 수집되는 데이터를 학습에 용이하게 정제하는 일련의 과정들을 일컫는 말이고, 여기서 사람의 직접적인 피드백을 모델에 반영하는 Human-in- the-loop 학습 방법론을 제시하는 점이 인상적이다 [35]. 또한, BlenderBot 2 의 강점이었던 인터넷 검색 모듈을 더 강화한 방법론도 다른 자매 논문에서 구체적으로 제시하고 있다 [39]. 가장 고무적인 점은 이 모든 데이터, 모델, 코드, 배포 방법론, UI 디자인까지 공개했다는 점으로 다양한 후속 연구가 기대된다.

대화에 특화된 LLM 외에도, OpenAI 의 GPT-3, Meta 의 OPT-175B [33], DeepMind의 Gopher [40], 그리고 다양한 기관의 연구자가 모여 협업한 BigScience 에서 발표한 BLOOM [41] 과 같이 GPT-3 와 유사한 크기를 가지는 모델들의 경우에 위 모델들과 달리 대화에 특화된 학습을 거치지 않았음에도 모두 어느 정도의 대화 능력을 갖추었다. 이는 Meena 논문에서 발표한 SSA 와 Perplexity 간의 높은 상관 관계로 인한 것으로 추정된다. 다만, Gopher 와 GPT-3 의 경우 대중들에게 공개된 모델이 아니어서 후속 연구에 제한사항이 있다.

이 섹션에서 다양한 종류의 대형 대화 모델들에 대해서 다루었는데, 다수의 연구 기관들에서 비슷한 시기에 여러가지로 공통적인 목적 의식을 가지고 연구를 진행한 것을 볼 수 있다. 그것들을 정리해보면 다음과 같다. 첫 째로, “안전한 대화”가 있고 이에 대해서는 다음 섹션에서 조금 더 다루고 있다. 두 번째로는 외부 지식을 활용한 대화 생성이다. 마지막 세 번째는, “좋은 대화란 무엇인가” 에 대한 고찰과 그것을 평가하는 다양한 방법론이다.

2.2. 챗봇 사용자의 안전을 고려한 설계 (Conversationally Safe Designs)

앞서 소개한 LamDA 와 BlenderBot 3에서 공통적으로 다루었던 주제 중 하나는 바로 챗봇 사용자의 안전을 고려한 설계다. LamDA 에서는 안정성과 관련된 규칙들을 제정하고 이를 기반으로 크라우드소싱을 진행해서 수집된 대화 데이터에 모델을 학습시켰다. 또한 안정성을 대화 모델의 주요 평가 지표로서 정의했다. BlenderBot 3에서는 LamDA 와 달리 SaFer Dialogues [37] 에 추가적으로 학습시키는 것 외에는 따로 외부적인 안정성 분류기를 BBF [42] 와 BAD [43] 데이터 등에 학습시켰다. 이는 대화 모델의 성능을 유지하기 위함도 있지만, 가장 중요한 점은 수집된 데이터가 Microsoft Tay [44] 의 경우처럼 Continual Learning 상황에서 모델을 오염시켜서 악영향을 끼치는 것을 막기 위한 이유가 크다 [36]. 이 외에도, 비슷한 문제 의식을 가진 몇 가지 연구들이 있었는데, 그 중 ProsocialDialogues [45] 라는 논문에서는 약 6만개의 안정성과 관련된 대화를 레이블링해서 공개했다. 또한, 이 대화 데이터를 바탕으로 학습된 Canary 라는 안정성 분류기 모듈을 만들었는데, 이 모듈을 통해 기존에 존재하는 대형 대화 모델들이 추가적인 학습 없이도 더 안전한 대화를 할 수 있다는 것을 보여주었다. 마지막으로, 최근 DeepMind 에서 Red-teaming (대항군) [46] 이라는 패러다임이 발표됐는데 Adversarial attack 과 비슷한 개념이다. 기존에 인공적으로 이루어지던 대화 모델에 대한 공격을 LLM 을 활용해서 (반)자동화 시킨 것이 가장 큰 기여점이다. 이 연구의 후속으로서 Anthropic 에서 발표한 논문은 이런 공격들을 체계적으로 데이터셋화 해서 공개했고 Red-teaming 이라는 개념을 더 명확하게 정리했다 [47].

3. 목적 지향형 대화 시스템 (Task-Oriented Dialogue Systems)

목적 지향형 대화 시스템의 연구 분야에서는 안타깝게도 상대적으로 LLM 과 같은 사전 학습이 아직까

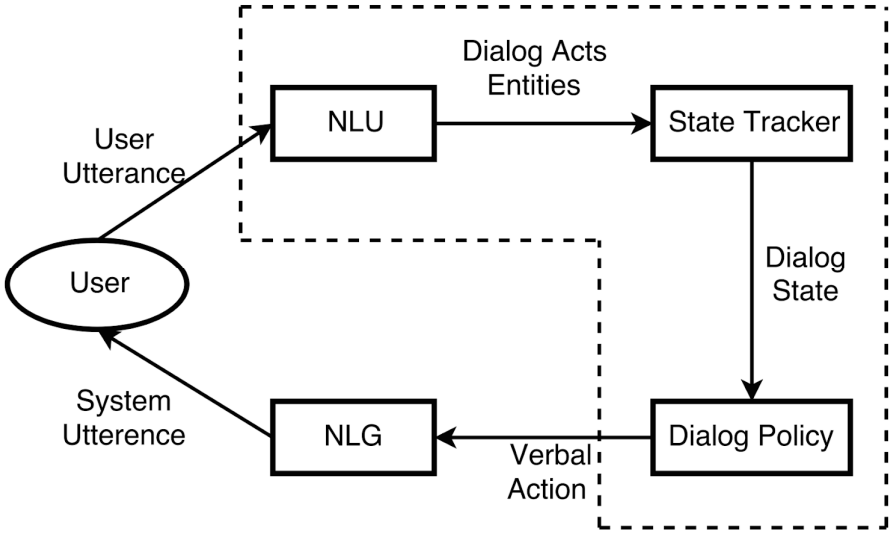

그림 2 목적 지향형 대화 시스템의 4가지 컴포넌트. 이용자의 전반적인 의도를 파악하는 NLU (Natural Language Understanding), 대화 상태를 추적하는 State Tracker, 모델이 다음으로 뱉어야 할 대화 Act 를 정하는 Dialog Policy, 그리고 정해진 대화 Act 를 자연어로 바꾸어 주는 NLG (Natural Language Generation) 컴포넌트로 이루어져 있다.
그림 출처: "Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning." T.Zhao and M.Eskenazi., 2016, SIGDIAL 2016

지는 쉽지 않은 상황이다. 이에 대한 원인으로는 1) 데이터 자체가 너무 수집하기 비싸고 어렵기 때문에 데이터의 양이 ODD 에 비해서는 극단적으로 적고, 2) 데이터의 특성상 대화 내용 뿐만 아니라 Database 와 같이 외부 모듈에서 정보를 가져와야 하는 경우가 많기 때문에 사전 학습으로만 해결하기에는 어렵다. 때문에, LLM의 발전 이후 이루어진 많은 연구가 그림 1에서 나오는 TOD 시스템의 각 컴포넌트를 높아진 LLM 성능을 통해 최적화하는 방향으로 이루어졌다. 이 섹션에서는 특히 대화 상태 추적 (Dialogue State Tracking) 이라는 TOD 시스템의 가장 핵심적인 기술이 어떻게 발전해가고 있는지 다루고 있고, LLM 을 이용한 종단간 (End-to-end) 모델들의 발전에 대해서도 간략히 소개한다.

3.1. 대화 상태 추적 (Dialogue State Tracking)

대형 언어 모델의 발전에 따라 TOD 연구 분야에서 매우 활발하게 이루어진 연구 방향은 이를 활용한 대화 상태 추적 기술의 발전이다 [48]. 우선, 공개되어 있는 다양한 대화 데이터셋들만 한데 모아서 다양한 TOD 학습 태스크들에 대하여 대규모 사전 학습을 한 TOD-BERT [49], SimpleTOD [50], SOLOIST [51], PPTOD [52] 와 같은 다양한 논문들이 발표되었다. 여기서 발표된 모델들은 기존의 대화 상태 추적에 필요한 다양한 휴리스틱들 없이도 비슷하게 높은 정확도를 이끌어냈기에 TOD를 위한 Foundation Model 로서 중요한 역할을 할 것으로 기대된다.

2019년에 발표된 TRADE [53] 논문에서 대화 상태 추적은 분류 기반 모델만 사용해야 한다는 관념을 타파하고 생성 기반 모델로 높은 성능과 확장성을 보여주었다. 특히 이는 BART [54], T5 [3] 와 같은 학계에서 용이하게 사용 가능한 사이즈와 준수한 성능을 지닌 사전 학습된 text-to-text 모델들의 공개와 맞물리면서 다양한 관련된 후속 연구들을 많이 이끌어 냈다. 이 중에서 많은 사람들이 관심을 가지고 연구한 방향은 바로 새로운 대화 도메인에 대해 더 적은 학습 데이터로 효율적으로 모델을 학습시키는 방법에 대한 것이다. 이러한 Few-shot & Zero-shot 학습을 위해서 크게 3가지 방향의 방법론들이 제시되었는데, 기본 골자는 대규모로 사전 학습된 LLM 들의 latent 하게 내장된 지식을 최대한 활용하자는 것이다. 첫 번째 방법은 새 도메인에 대해 대화 상태 추적을 할 때 본 적 없는 Slot 과 Value 에 대한 자연어로 된 설명들을 T5 같은 모델에 Prompt 로서 전달하는 방법이다 [55,56]. 이 방법의 단점으로는 새로 적용해야 하는 도메인의 성질이 기존에 학습된 것들과 비슷해야 잘 작동한다는 것이다. 그래서 나온 두 번째 방법은 Question Answering 형식으로 LLM 에 각 슬롯에 대해 직접적으로 질문을 하는 방식이다 [57,58]. 마지막 세 번째

방법은 앞선 QA 형식의 비효율성을 꼬집으며 대화 상태 추적을 대화 요약이라는 형태로 바꾸면서 더 좋은 성능과 효율성을 보여주었다 [59]. 하지만, 아직까지는 이러한 Few-shot & Zero-shot Dialogue State Tracking 방법론들이 산업에서 실제로 사용 가능한 수준으로 성능이 나오려면 훨씬 더 많은 발전이 필요한 것으로 보인다.

3.2. 종단간 목적 지향형 대화 시스템 (End-to-End Task-oriented Dialogue Systems)

종단간 TOD 모델이란 (End-to-end TOD model) 일반적인 TOD 시스템처럼 컴포넌트 형태로 파이프라인을 구축하는 것이 아닌, 하나의 모델이 한 번의 계산으로 대화를 끝까지 생성해내는 것을 말한다 [60,61]. 대규모로 사전 학습된 LLM 들이 부각되기 시작하면서 이들의 학습 방법론을 채용해서 종단간 대화 모델을 학습시킨 연구들이 몇몇 등장했다. 그 중에서 2020년에 발표된 SimpleTOD [50] 와 SOLOIST [51] 는 그림 1의 각 컴포넌트의 입출력 데이터 형태를 모두 텍스트 형태로 치환해서 GPT-2 형태의 언어 모델을 학습시켰고 MultiWoZ [62] 데이터에서 높은 성능을 기록했다[5]. 두 모델의 가장 큰 차이는 사전 학습 데이터의 유무로 SOLOIST 는 7개의 추가적인 대화 데이터에 사전 학습을 시킴으로써 MultiWoZ 데이터에서 높은 Few-shot 학습 성능을 보였다. 그리고 비슷한 시기에 발표된 MinTL [63] 에서는 대화 데이터에 대한 특별한 사전 학습 없이도 높은 Few-shot 학습 성능을 보였다. 이후 2022년에 발표된 GALAXY [64] 와 PPTOD [52] 에서는 훨씬 더 발전된 성능의 사전 학습된 TOD 대화 모델들이 공개되었다. Bordes et al. [60] 의 연구 이후로 종단간 TOD 모델의 연구에 많은 진척이 없었고 대부분 대화 상태 추적에 관심이 쏠려 있었다. 대규모로 사전 학습된 대형 언어 모델들의 발전이 촉매가 되어서 이처럼 종단간 TOD 모델의 연구가 빠른 속도로 이루어지는 것은 매우 고무적인 일이고 앞으로도 지속되길 바란다.

4. 토의 및 결론

본 논문에서는, 대규모 사전 학습 대화 모델의 거듭된 승전 이후 두 가지 종류의 대화 시스템에 관한 연구가 어떠한 방식으로 발전하고 있는지 다루었다. 오픈 도메인 대화의 경우 데이터가 매우 많은 편에 속하기 때문에 직접적으로 LLM 의 학습 방식을 채용해서 대형 대화 모델을 만드는 방향의 연구들이 많이 공개되었고, 목적 지향형 대화의 경우에는 사전 학습된 LLM 을 직접적으로 사용해서 더 효율적이고 효과적인 학습 방법들에 관한 연구들이 많이 이루어졌다.

다만, 개인적으로 안타까웠던 점은 세 가지다. 첫째는, 아직까지 상용화될 수 있는 수준의 모델들이 공개되지 않은 것이다. 둘째는, 본 논문에서 다룬 눈부신 발전을 이룩한 대형 대화 모델들이 전부 Google, Meta, Microsoft와 같이 초거대 기업들이 연구하고 발표한 것들이라는 점이다. 마지막으로 셋째는, 아직 대규모 사전 학습의 스케일이 목적 지향형 대화에는 많이 적용되지 못했다는 점이다. 추후에 훨씬 더 많은 기관들에서 다양한 관점으로 연구를 진행함으로써 대화 모델 연구가 상용화를 이루기까지 현재의 동력을 잃지 않고 발전해 나가기를 바라며 논문을 맺는다.

5) 다만, SimpleTOD 모델의 DST 성능의 경우 재현성 이슈가 있는 것으로 알려져 있다.

5. 참고 문헌

[1] Brown, Tom, et al. “Language models are few-shot learners.” Advances in neural information processing systems 33 (2020): 1877-1901.

[2] Rae, Jack W., et al. “Scaling language models: Methods, analysis & insights from training gopher.” arXiv preprint arXiv:2112.11446 (2021).

[3] Raffel, Colin, et al. “Exploring the limits of transfer learning with a unified text-to-text transformer.” J. Mach. Learn. Res. 21.140 (2020): 1-67.

[4] Gao, Jianfeng, Michel Galley, and Lihong Li. “Neural approaches to conversational AI.” The 41st international ACM SIGIR conference on research & development in information retrieval. 2018.

[5] Chen, Hongshen, et al. “A survey on dialogue systems: Recent advances and new frontiers.” Acm Sigkdd Explorations Newsletter 19.2 (2017): 25-35.

[6] Ni, Jinjie, et al. “Recent advances in deep learning based dialogue systems: A systematic survey.” Artificial Intelligence Review (2022): 1-101.

[7] Radford, Alec, et al. “Improving language understanding by generative pre-training.” OpenAI blog (2018).

[8] Radford, Alec, et al. “Language models are unsupervised multitask learners.” OpenAI blog (2019).

[9] Zhang, Yizhe, et al. “DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics:

System Demonstrations. 2020.
[10] Zhou, Li, et al. “The design and implementation of xiaoice, an empathetic social chatbot.” Computational Linguistics 46.1 (2020): 53-93.
[11] Rishi Bommasani, et al. “On the opportunities and risks of foundation models.” arXiv preprint arXiv:2108.07258 (2021).
[12] Wolf, Thomas, et al. “Transfertransfo: A transfer learning approach for neural network based conversational agents.” arXiv preprint arXiv:1901.08149 (2019).
[13] Peng, Baolin, et al. “GODEL: Large-Scale Pre-Training for Goal-Directed Dialog.” arXiv preprint arXiv:2206.11309 (2022).
[14] Nguyen, Tri, et al. “MS MARCO: A human generated machine reading comprehension dataset.” CoCo@ NIPs. 2016.
[15] Khashabi, Daniel, et al. “UNIFIEDQA: Crossing Format Boundaries with a Single QA System.” Findings of the Association for Computational Linguistics: EMNLP 2020. 2020.
[16] Galley, Michel, et al. “Grounded response generation task at dstc7.” AAAI Dialog System Technology Challenges Workshop. 2019.
[17] Rastogi, Abhinav, et al. “Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05. 2020.
[18] Ben Wang and Aran Komatsuzaki. 2021. GPTJ-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax.
[19] Vinyals, Oriol, and Quoc Le. “A neural conversational model.” arXiv preprint arXiv:1506.05869 (2015).
[20] Adiwardana, Daniel, et al. “Towards a human-like open-domain chatbot.” arXiv preprint arXiv:2001.09977 (2020).
[21] So, David, Quoc Le, and Chen Liang. “The evolved transformer.” International Conference on Machine Learning. PMLR, 2019.
[22] Thoppilan, Romal, et al. “Lamda: Language models for dialog applications.” arXiv preprint arXiv:2201.08239 (2022).
[23] Roller, Stephen, et al. “Recipes for Building an Open-Domain Chatbot.” Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021.
[24] Rashkin, Hannah, et al. “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset.” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
[25] Zhang, Saizheng, et al. “Personalizing Dialogue Agents: I have a dog, do you have pets too?.” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
[26] Dinan, Emily, et al. “Wizard of Wikipedia: Knowledge-Powered Conversational Agents.” International Conference on Learning Representations. 2018.
[27] Smith, Eric Michael, et al. “Can You Put it All Together: Evaluating Conversational Agents’ Ability to Blend Skills.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.
[28] Li, Margaret, Jason Weston, and Stephen Roller. “Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons.” arXiv preprint arXiv:1909.03087 (2019).
[29] Komeili, Mojtaba, Kurt Shuster, and Jason Weston. “Internet-Augmented Dialogue Generation.” Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.
[30] Xu, Jing, Arthur Szlam, and Jason Weston. “Beyond Goldfish Memory: Long-Term Open-Domain Conversation.” Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.
[31] Lewis, Patrick, et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks.” Advances in Neural Information Processing Systems 33 (2020): 9459-9474.
[32] Izacard, Gautier, and Edouard Grave. “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering.” EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2021.
[33] Shuster, Kurt, et al. “Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.” arXiv preprint arXiv:2208.03188 (2022).
[34] Zhang, Susan, et al. “Opt: Open pre-trained transformer language models.” arXiv preprint arXiv:2205.01068 (2022).

[35] Xu, Jing, et al. “Learning New Skills after Deployment: Improving open-domain internet-driven dialogue with human feedback.” arXiv preprint arXiv:2208.03270 (2022).

[36] Ju, Da, et al. “Learning from data in the mixed adversarial non-adversarial case: Finding the helpers and ignoring the trolls.” arXiv preprint arXiv:2208.03295 (2022).

[37] Ung, Megan, Jing Xu, and Y-Lan Boureau. “SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures.” Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

[38] Xu, Jing, et al. “Recipes for safety in open-domain chatbots.” arXiv preprint arXiv:2010.07079 (2020).

[39] Shuster, Kurt, et al. “Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion.” arXiv preprint arXiv:2203.13224 (2022).

[40] Rae, Jack W., et al. “Scaling language models: Methods, analysis & insights from training gopher.” arXiv preprint arXiv:2112.11446 (2021).

[41] Scao, Teven Le, et al. “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.” arXiv preprint arXiv:2211.05100 (2022).

[42] Dinan, Emily, et al. “Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack.” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[43] Xu, Jing, et al. “Bot-adversarial dialogue for safe conversational agents.” Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

[44] Ernest Davis. 2016. Ai amusements: the Tragic Tale of Tay the Chatbot. AI Matters, 2(4):20–24.

[45] Kim, Hyunwoo, et al. “ProsocialDialog: A Prosocial Backbone for Conversational Agents.” arXiv preprint arXiv:2205.12688 (2022).

[46] Perez, Ethan, et al. “Red teaming language models with language models.” arXiv preprint arXiv:2202.03286 (2022).

[47] Ganguli, Deep, et al. “Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned.” arXiv preprint arXiv:2209.07858 (2022).

[48] Jacqmin, Léo, Lina M. Rojas Barahona, and Benoit Favre. ““Do you follow me?”: A Survey of Recent Approaches in Dialogue State Tracking.” Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2022.

[49] Wu, Chien-Sheng, et al. “TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue.” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[50] Hosseini-Asl, Ehsan, et al. “A simple language model for task-oriented dialogue.” Advances in Neural Information Processing Systems 33 (2020): 20179-20191.

[51] Peng, Baolin, et al. “Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model.” arXiv preprint arXiv:2005.05298 (2020).

[52] Su, Yixuan, et al. “Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System.” Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

[53] Wu, Chien-Sheng, et al. “Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems.” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[54] Lewis, Mike, et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[55] Lin, Zhaojiang, et al. “Leveraging Slot Descriptions for Zero-Shot Cross-Domain Dialogue StateTracking.” Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

[56] Zhao, Jeffrey, et al. “Description-Driven Task-Oriented Dialog Modeling.” arXiv preprint arXiv:2201.08904 (2022).

[57] Lin, Zhaojiang, et al. “Zero-Shot Dialogue State Tracking via Cross-Task Transfer.” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

[58] Gao, Shuyang, et al. “From Machine Reading

Comprehension to Dialogue State Tracking: Bridging the Gap.” Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI. 2020.

[59] Shin, Jamin, et al. “Dialogue Summaries as Dialogue States (DS2), Template-Guided Summarization for Few-shot Dialogue State Tracking.” Findings of the Association for Computational Linguistics: ACL 2022. 2022.

[60] Bordes, Antoine, Y-Lan Boureau, and Jason Weston. “Learning end-to-end goal-oriented dialog.” arXiv preprint arXiv:1605.07683 (2016).

[61] Wen, Tsung-Hsien, et al. “A Network-based End-to-End Trainable Task-oriented Dialogue System.” Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. 2017.

[62] Budzianowski, Paweł, et al. “MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling.” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[63] Lin, Zhaojiang, et al. “MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems.” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[64] He, Wanwei, et al. “Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 10. 2022.

약 력

신 재 민

2017 홍콩과기대 (HKUST) Computer Science 졸업 (학사)
2020 홍콩과기대 (HKUST) Electrical and Computer Engineering 졸업 (석사, MPhil)
2020 Amazon Alexa AI, Research Intern
2020~2022 Riiid AI Research, Research Manager
2022~현재 NAVER AI Lab, Research Scientist
관심분야: Large Language Models, Dialogue Systems, Trustworthy AI
Email: jamin.shin@navercorp.com