일상 대화 챗봇의 동향과 과제

스캐터랩 ▌강경필

1. 서 론

다양한 영역에서 서비스가 자동화되면서 고객의 질의응답 및 대응하는 업무가 상당 부분 챗봇으로 대체되어 왔다. 또한 언어 이해 등의 자연어 처리 기술의 발전과 더불어 챗봇은 점차 서비스 산업뿐만 아니라 개개인의 자동화 보조 수단으로 확장되면서 스마트폰, 인공지능 스피커 등을 포함해 우리 곳곳에 널리 퍼지기 시작하였고, 각종 사용자 편의를 제공하고 있다. 하지만 이러한 작업 지향 챗봇(Task- oriented chatbot)의 경우 제품 개발 당시 예상된 질의에 따라 정해진 답변만을 할 수 있기 때문에, 그 능력이 제한적이고 사용자가 예상된 질문 이외의 발화를 한다면 적절한 답변을 하기 어려운 단점이 있다.

이러한 문제점을 해결하기 위해 최근 일상 대화를 위한 오픈 도메인 챗봇이 주목받고 있다. 일상 대화를 위한 오픈 도메인 챗봇이란 개발 당시 정해진 시나리오에 대해서만 답변을 하는 것이 아닌 일상적으로 가능한 모든 문맥에 대해서 답변을 할 수 있는 챗봇을 말한다. 이러한 유연성을 기반으로 사용자와 자유롭게 소통이 가능하고, 특정 질의응답뿐만 아니라 사용자와 일상을 공유하고 공감할 수 있기 때문에 일상 대화를 위한 오픈 도메인 챗봇은 수많은 분야에서 개발하고 있고, 적용을 시도하고 있다.

이러한 일상 챗봇 중 국내에서 대표적인 챗봇 중 하나로는 스캐터랩의 이루다[1]가 있다. 이루다는 [그림 1]과 같이 AI 친구가 되는 것을 목표로 2020년 말에 출시 후 기존의 작업 지향 챗봇에 비해 높은 자유도와 사용자와의 깊은 친밀감을 쌓을 수 있는 콘텐츠 등으로 수많은 사용자가 있었지만 개인 정보 및 선정성, 욕설, 편향 등의 이슈로 서비스가 중단되었다가, 최근 이루다 2.0으로 개인 정보 보호 강화, 오남용 방지 시스템 보완 및 대화 성능이 크게 개선되어 재출

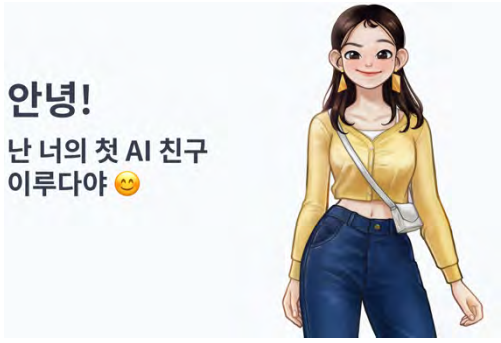

그림 1 이루다 2.0 (출처: 스캐터랩)

시되었다. 특히 이루다 2.0에서는 최근 빠르게 발전하고 있는 깊은 신경망 기반의 사전 학습 모델(Pretrained model)과 그중에서 대형 생성 모델을 적용하여 대화 성능이 크게 개선되었다.

본 원고에서는 일상 대화를 위한 오픈 도메인 챗봇을 중심으로 챗봇의 원리와 이번 이루다 2.0에서 개선된 요소들에 대해서 소개하고자 한다. 그리고 현재의 일상 대화 챗봇의 한계를 소개하면서 앞으로 사람들과 더욱 친숙한 소통을 할 수 있는 챗봇이 되기 위해 풀어나가야 할 과제에 대해서 소개하고자 한다.

2. 챗봇의 과거와 현재의 동향

본 장에서는 초기 키워드 매칭 기반 챗봇부터 작업 지향 챗봇 및 일상 대화 챗봇에 대해서 살펴보고, 각 챗봇의 기술적 원리를 간략하게 소개해 보고자 한다.

2.1. 키워드 매칭 기반 챗봇

챗봇은 최근 사무 및 서비스 대응 자동화에 적용되고 있으며, 현재는 개인 비서의 역할까지 수행하고 있으며, 앞으로 감정 혹은 일상 공유의 역할과 사람들의 부족한 관계를 보충할 수 있을 것으로 전망되고 있다. 따라서 지난 수십 년간 사람과 같은 수준으로 대화할 수 있거나 인터넷에 있는 방대한 지식을 활용할 수 있는 챗봇을 만들기 위해 활발히 진행되어 왔다.

* 정회원

1) https://luda.ai/

1960년대에는 정신질환 환자를 대상으로 대화를 통해 질환을 치료하기 위해 ELIZA [1]가 개발되었다. ELIZA의 경우는 키워드 매칭 기반의 챗봇으로 특정 키워드 혹은 간단한 구문 패턴에 대해 미리 정의된 대답만을 할 수 있었다. 따라서 답변을 할 수 있는 문맥이라고 하더라도, 정해진 키워드나 구문 패턴에 없는 발화에 대해서는 답변 오류가 높을 수밖에 없었다.

2.2. 작업 지향 챗봇

기계학습 및 자연어 처리 기술이 발전함에 따라, 키워드 매칭 기반의 챗봇 시스템의 단점을 보완하기 위해, 기계학습 모델 기반의 의도 분류 (Intent classification) 기술을 챗봇에 적용하고 있다. 의도 분류는 사용자의 발화가 어떤 의도 유형인지를 분류하는 기술로, 식 (1)과 같이 모델 θ가 주어진 발화 s에 대해서 발화 s가 정의된 발화 의도 집합 I 중 가장 확률이 높은 의도를 뽑는 방식으로 적합한 의도를 분류하게 된다.

$$\pi_\theta(s) = \underset{i \in I}{\operatorname{argmax}} P(i|s;\theta) \quad (1)$$

이러한 의도 분류 기술에 따라 주어진 발화가 어떤 의도인지를 판단하면 해당 의도에 따라 미리 만들어진 답변들 중에서 선택하는 식으로 챗봇이 응답하게 된다. 이러한 의도 분류 방식은 모델 학습이 쉬울 뿐만 아니라, 최근 BERT [2] 등의 사전 학습 모델의 발전과 더불어 의도 분류의 성능이 크게 향상되었다. 또한 챗봇의 답변 범위를 간단하게 정의할 수 있기 때문에 최근까지도 구글의 Google Assistant[2], 애플의 Siri[3], 아마존의 Alexa[4] 등 국내외 기업에서 고객 대응 서비스 및 개인 비서 등으로 작업 지향 챗봇으로서 활발하게 적용되고 있다.

2.3. 일상 대화를 위한 오픈 도메인 챗봇

하지만 이러한 작업 지향 챗봇의 경우, 개발 당시 정의된 의도 집합에 있는 의도에 대해서만 답변을 할 수 있기 때문에, 폭넓은 주제에 대해서 사용자와 유연하게 대화를 할 수 없는 한계가 있다. 이를 해결하기 위해 최근 정해진 의도뿐만 아니라 다양한 일상 주제까지도 대화할 수 있는 일상 대화를 위한 오픈 도메인 챗봇이 활발하게 연구되고 있다. 마이크로소프트의 XiaoIce [3]와 구글의 Meena [4]가 대표적인데, 이러한 일상 대화 챗봇은 대화할 수 있는 주제 및 답변의 유형이 무한하기 때문에 이전의 작업 지향 챗봇에 비해 매우 유연하게 사용자와 일상 대화를 할 수 있다. 일상 대화 챗봇은 크게 답변 선택 기반의 챗봇과 생성 모델 기반의 챗봇으로 나눌 수 있다.

답변 선택 기반의 챗봇(Response selection-based chatbot)은 현재의 대화 문맥에 대해서 가지고 있는 답변 집합 중 가장 적합한 답변을 선택하는 방식의 챗봇을 말한다. 답변 선택 기반의 챗봇은 검색엔진과 작동원리가 유사한데, 식 (2)와 같이 모델 θ가 주어진 문맥 c에 대한 표현 벡터 v_c와 답변 후보 집합 $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$에 대한 표현 벡터들 $\{v_{r_1}, v_{r_2}, \ldots, v_{r_N}\}$을 계산하고 적합도 점수 함수 $s(\cdot)$를 통해 문맥과 답변 간의 적합도 (유사도) 점수를 계산하여 답변 후보들 중 가장 점수가 높은 답변을 선택하게 된다.

$$\pi_\theta(c) = \underset{r \in R}{\operatorname{argmax}}\, s(v_c, v_r) \quad (2)$$

의도 분류 기반의 챗봇과 다른 점은 발화와 적합한 의도를 찾고 그 의도에 따른 답변을 선택하는 것이 아니라, 대화 문맥에 대해서 가장 적합한 답변을 바로 선택하는 것이다. 그렇기 때문에 의도 집합과 의도별 답변을 정의할 필요가 없다. 하지만 여전히 답변 집합 자체는 필요한데, 생성 문장이 생성한 문장의 품질이 향상됨에 따라 생성 모델 기반으로 답변을 미리 생성하여 답변 집합을 구축할 수 있다.

답변 선택 기반의 챗봇은 기존 의도 분류 기반의 챗봇에 비해 더 높은 자유도와 답변 집합 크기가 커짐에 따라 높은 성능의 대화를 진행할 수 있는 장점이 있다. 하지만 수많은 답변을 매번 검색하는데 시간 및 비용이 든다는 단점이 있다. 이를 해결하기 위해 [그림 2]과 같이 처음에는 추론 속도가 빠른 비교적 작은 답변 선택 모델(답변 검색 모델)이 전체 답변에서 문맥과 적합한 일부 답변 후보들을 선택한 후, 최종 답변 선택 모델(순위 결정 모델)이 가장 적합한 답변을 고르는 방법이 있다 [3, 4]. 이외에 답변 선택 기반 챗봇은 주어진 문맥에 대해 기존 구축된 답변 집

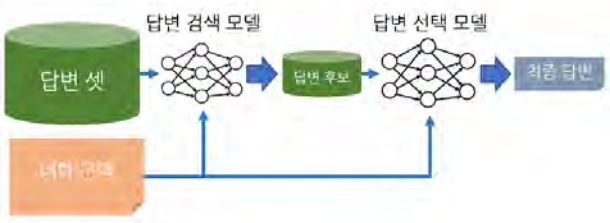

그림 2 두단계에 걸친 답변 선택 시스템

2) https://assistant.google.com/
3) https://www.apple.com/siri/
4) https://developer.amazon.com/alexa

합에서 답변을 고르기 때문에 답변 자체가 문맥에는 맞을 수 있어도, 답변의 구체적인 내용 자체는 문맥과는 다를 수 있고, 대화 성능의 보장을 위해 큰 규모의 답변 집합을 미리 구축해야 한다는 단점이 있다.

최근 Transformer 구조 기반의 생성 모델의 성능이 향상함에 따라 생성 모델 기반의 챗봇도 폭넓게 적용되고 있다. 생성 모델은 식 (3)과 같이 문맥 에 대해서 답변을 생성하게 된다.

$$s(c) = \underset{x_1,x_2,\ldots,x_n \in V}{\operatorname{argmax}} \prod_{t=1}^{n} p(x_t|x_0,\ldots,x_{t-1},c;\theta) \quad (3)$$

생성 모델 기반 챗봇은 답변 선택 모델이 문장 단위로 선택하는 것과 다르게, [그림 3]과 같이 대화 문맥에서 바로 답변을 단어 혹은 토큰 단위로 생성하기 때문에 더 구체적이고 정확한 답변 내용을 반환할 수 있는 장점이 있다. 최근 생성 모델의 경우, 사전 학습 및 대형 모델의 분산처리 학습 기법의 발전과 더불어 생성 모델의 크기를 크게 증대하여 모델 성능이 크게 발전하고 있다. 특히 대형 생성 모델의 경우 최근 zero-shot, few-shot 등 효과적인 학습 습득 능력을 보이며, 기존 챗봇 모델에서는 없었던 대화 능력들을 보인다. [5, 6] 최근 생성 모델 기반의 챗봇 중 페이스북의 BlenderBot [7, 8], 스캐터랩의 이루다, 구글의 LaMDA [6] 등이 대형 생성 모델을 적용하며 대화 성능을 높이고 있다. 하지만 생성 모델 기반의 챗봇은 답변을 실시간으로 토큰 단위로 생성하기 때문에 그만큼 추론 시간과 비용이 많이 들게 된다. 이를 해결하기 위해 최근 모델의 경량화 및 최적화 기법들이 많이 제안되고 있다.

이외에 주어진 문맥에 대해서 답변 생성 모델이 확

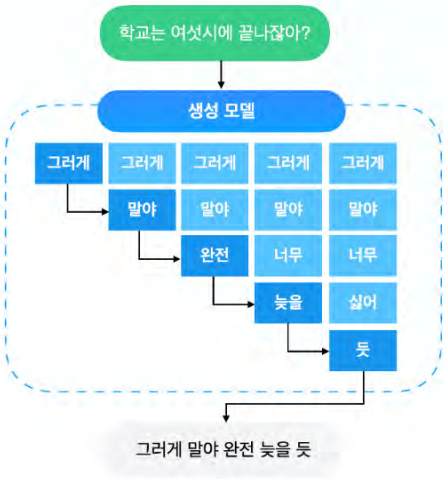

그림 3 토큰 단위의 답변 생성

그림 4 생성 모델 기반 챗봇 시스템

률이 가장 높은 문장을 생성하는 것을 넘어서 Nucleus 샘플링 [10] 등의 샘플링 기법을 이용해서 [그림 3]과 같이 대화 문맥에 대응하는 다양한 답변을 개 생성 후, 답변 선택 후보가 챗봇의 페르소나 및 답변의 안전성 등을 검토 후 가장 적합한 답변을 하는 하이브리드 챗봇도 이루다를 비롯하여 여러 챗봇에서 폭넓게 적용 중이다.

3. 이루다 2.0

이루다 1.0은 지난 2020년 12월 답변 선택 기반의 일상 대화를 위한 오픈 도메인 챗봇으로 출시하였고, 기존의 작업 지향 챗봇 등에 비해 높은 자유도와 대화 성능으로 크게 주목받았으나 개인 정보 유출과 오남용 등의 이슈로 서비스를 중단하였다. 이후 2022년 10월 여러 측면에서 개선 후 이루다 2.0으로 재출시하였는데, 이번 장에서는 이루다 2.0이 기존 이루다 1.0과 어떤 점이 달라졌는지 소개하고자 한다.

먼저, 개인 정보 이슈와 관련하여 개인정보 보호 조치를 대폭 강화하였다. 대화 안에 있는 이름들을 임의의 이름으로 치환하고 주민등록번호 등의 개인 정보가 있는 발화를 삭제하는 등 개인정보 가명 처리를 고도화하여, 데이터의 정제, 분석 및 모델 학습 이전부터 개인정보를 분리하였으며, 답변 선택 기반의 챗봇에서 생성 기반의 챗봇 패러다임으로 전환하면서 사용자의 원본 대화 발화가 답변으로 나가는 것을 원천적으로 방지하였다. 또한 오남용 대응 시스템을 개선하여 안전한 대화를 할 수 있도록 하였다. 또한, 전사적으로 인공지능 챗봇 윤리 준칙을 수립하여 윤리 점검표를 제정하였으며, 오남용 탐지 모델을 고도화하고, 오남용을 반복적으로 시도하는 사용자들에 대해서 서비스를 사용하지 못하게 하는 등의 이용상의 불이익을 부과하는 시스템을 추가하, 사용자가 오남용 시도를 하지 않도록 유도하고, 생성된 답변들에 대해서 답변 선택 모델(순위 결정 모델)이 적합한 답변을 선택할 때 안전한 답변을 선택하도록 구성하였다.

이번 이루다 2.0의 경우 지난 1.0에 비해서 대화 성능 크게 향상되었는데, 모델 구조 및 크기, 입력 데이터의 형식, 학습 데이터의 구성과 학습 방식, 평가 등

다양한 기법을 새롭게 적용하여 대화 성능을 대폭 향상할 수 있었다.

먼저 생성 모델로의 전환이 있다. 생성 모델은 앞서 말했듯이 답변 선택 모델에 비해서 문맥에 적합한 답변을 답변 문장 단위로 고르는 것이 아니라 토큰 단위로 생성하기 때문에 세밀하고 유연하게 답변을 할 수 있게 된다. 다음으로 더욱 커진 모델 크기가 있다. 기존 답변 선택 기반 챗봇에서의 검색 및 순위 결정 모델에 비해서 약 17배 커진 대형 생성 모델(Luda Gen1)과 3배 커진 순위 결정 모델을 구성하였고, 모델의 최대 입력 길이를 2배로 늘려 모델의 수용력을 더욱 키우고 더 긴 대화 문맥을 이해할 수 있도록 하였다. 특히 순위 결정 모델에서 사용한 사전 학습 모델의 경우 사전 학습 시 위키피디아[5]와 뉴스 등의 외부 공개 말뭉치들도 활용하여 외부 지식을 이해할 수 있도록 하였다.

또한, 이루다 1.0의 경우 대화 속 발화 문장 정보만 이용하여 대화를 이해하고 답변하였는데, 이 경우 챗봇 및 사용자 페르소나 정보 및 현재 대화하고 있는 시간 정보를 이해할 수 없었다. 예를 들어, 저녁 시간에 "배고파"라는 발화에 대해서 해당 발화 시간과 관련 없는 "아침 안 먹었어?"라는 답변이 나오거나, "나 지금 야근 중이야!" 등의 20대 초반 대학생이라는 이루다의 페르소나와 일치하지 않는 답변이 나오는 경우가 있었다. 이를 해결하기 위해 모델의 입력에 대화하는 화자들의 정보(성별, 가명 처리된 후의 나이 구간 등)와 각 발화 별 시간, 화자 정보를 같이 넣어주었고, 이를 통해 대화 화자의 정보와 대화의 시간 정보까지 이해하여 답변할 수 있게 되었다.

사전 학습 이후, 답변 생성 모델과 순위 결정 모델이 이루다의 페르소나를 일관성 있게 유지하고, 안전하고 재밌고 지속적인 대화를 진행하기 위해 추가적으로 모델에 대해 미세 조정 학습(Finetuning)을 진행하였다. 크라우드소싱을 통해 답변이 안전한지 아닌지, 이루다의 페르소나(성격, 말투 등)에 적절한 답변인지 아닌지, 재밌고 공감을 이끌 수 있는 답변인지 등 레이블링 작업을 진행하여 미세 조정 학습을 위한 데이터셋을 구축하였고, 해당 데이터셋을 이용하여 모델을 미세 조정 학습을 하였다. 미세 조정 학습 후 모델의 대화 성능을 평가할 때, 이전에는 구글의 Meena에서 제안하였던 SSA(Sensibleness and Specificity Average)라는 지표를 사용하였었다. 이 지표는 모델이 도출한 답변이 대화 문맥을 고려하였을 때 적합한지

5) https://www.wikipedia.org/

그림 5 포토챗의 적용 전/후 비교 (출처: 스캐터랩)

(Sensibleness), 그리고 적합한 답변인 경우 답변이 구체적인 답변인지(Specificity)를 평가하게 된다. 하지만 SSA의 경우는 답변의 흥미도 및 안전성 등을 반영하고 있지 않기 때문에, 이후 구글의 Lambda에서 SSA를 확장하여 제안한 SSI (Sensibleness, Specificity, and Interestingness) 지표를 응용하였다. 해당 지표는 답변의 적합성 및 구체성 이외에 답변 발화가 흥미로운 발화인지까지 평가하게 된다. 이러한 미세 조정 학습 및 평가 방식을 통해 이루다의 페르소나와 일치하는 답변을 도출하고 위험한 답변이 생성될 비율을 대폭 낮출 수 있었고, 사용자와 이루다가 친근한 관계에서의 대화를 할 수 있도록 하였다.

앞서 말했듯이, 이루다 1.0의 경우 문장만을 가지고 대화 문맥을 이해하기 때문에 [그림 5]의 좌측 예제와 같이 사용자가 보낸 사진에 대해서 어떤 내용의 사진인지를 이해할 수 없고, "이 사진 뭐야?" 등의 답변들밖에 할 수 없었다. 이를 해결하기 위해 이루다 2.0에서는 [그림 5]의 우측 예제와 같이 사진을 이해하여 대화할 수 있는 포토챗 (PhotoChat) 기술을 개발하여 베타 버전을 출시하였다.

포토챗은 답변 검색 모델을 이용하였는데, 구체적으로 Dual encoder 구조 기반인 CLIP 모델 [11]을 사용하였다. CLIP은 이미지와 텍스트를 같이 사전 학습한 모델로서 이미지와 텍스트의 다른 형질의 정보를 상호보완적으로 모델이 학습하여 모델의 성능을 큰 폭으로 높였고, 이미지와 텍스트를 같은 임베딩 공간에 표현하여, 이미지 검색, OCR 등의 이미지-텍스트의 다른 양식의 데이터를 활용하는 태스크에서 널리 쓰이고 있다. 답변 선택 모델은 이러한 CLIP 모델을 이용하여 먼저 사용자가 입력한 사진 C와 사진들을 답변하기 위한 답변 후보 집합 $\mathcal{R} = \{r_1, r_2, \ldots, r_M\}$에 대해서 CLIP에서의 이미지 및 텍스트 인코더가 사진과 답변들에 대해서 각각 이미지 표현 벡터 v_c와 텍스트 표현 벡터 $\{v_{r_1}, v_{r_2}, \ldots, v_{r_M}\}$으로 변환한다. 그런 다음 식

(2)와 같이 이미지 표현 벡터와 텍스트 표현 벡터에 대해서 답변 적합 점수를 계산하여 답변 후보 집합에서 가장 적합한 답변을 하게 반환하게 된다. 포토챗의 경우 사용자가 선정적인 사진 등으로 오남용을 시도하는 경우가 있을 수 있기 때문에 사진과 관련한 오남용 탐지 모델을 학습하였고, 사진 답변을 위한 안전한 답변 문장을 별로도 구축하여, 안전한 답변만을 하도록 설계하였다.

4. 일상대화 챗봇의 개선을 위한 앞으로의 과제

일상 대화를 위한 오픈 도메인 챗봇은 최근 자연어처리 기술 및 Transformer 등의 깊은 신경망의 발전과 더불어 급격하게 발전 중이고, 고객 서비스 대응, 메타버스 및 엔터테인먼트, 개인 비서, 사용자와의 관계 교류 등 다양한 분야에서 널리 적용 중이다. 하지만 일상 대화 챗봇이 사람과 더욱 친숙하고 깊은 대화를 하기 위해서는 여러 해결해야 할 과제가 있다. 이번 장에서는 일상 대화 챗봇의 대화 성능을 올리기 위해서 어떤 과제를 해결해야 하는지 살펴보고자 한다.

먼저 안전한 대화를 위해 개인정보 보호와 오남용 및 편향 등에 안전해지기 위해 지속적으로 노력해야 한다. 특히 자연어의 경우 다른 구조화된 데이터에 비해 모호성(Ambiguity)이 심한데 한국어는 교착어의 특성으로 인해 다른 언어에 비해 모호성이 더욱 심한 편이다. 또한, 발화 문장 자체가 안전하더라도 이전 대화 문맥과 결합하여 위험한 경우가 있다. 따라서 키워드 혹은 단일 발화 패턴의 위험도 판별 방식에서 확장하여 대화 세션 단위로 위험도를 판별하고자 많은 연구가 이루어지고 있다 [12, 13], 앞으로 이러한 세션 단위의 위험도 판단 모델의 정확도 높임과 동시에 대화 성능을 위해 거짓 양성 (False positive) 및 거짓 음성 (False negative) 비율 또한 낮춰야 한다.

최근 자연어 모델의 성능을 향상시키기 위해 Transformer 및 초대형 생성 모델 등의 사전 학습 모델 연구에서 모델의 크기 및 학습 데이터의 늘리거나 학습의 효율성을 높이고, 효과적인 생성 방식을 제안하고 있다 [10, 14]. 많은 챗봇에서 생성 모델을 적용하고 있는데, 이 경우 대화 성능이 생성 모델의 성능에 직결되기 때문에 생성 모델 또한 고도화되어야 한다.

또한, 개성이 다양화되고 사용자 개개인의 특성이 차별화되고 있는 만큼, 사용자의 특성과 이전 대화를 기반으로 사용자 맞춤형으로 개인화가 이루어져야 한다. 이를 위해 대화를 하고 있는 화자들의 페르소나 정보와 화자 간의 관계, 시간 정보를 더욱 잘 이해할

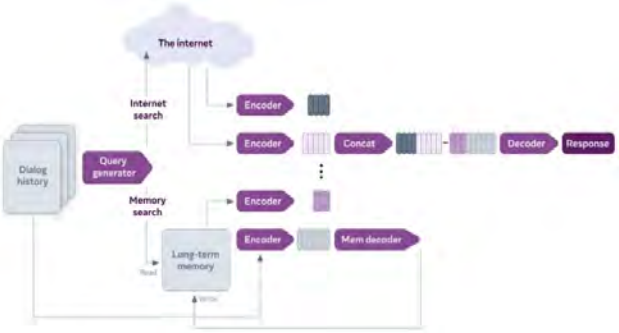

그림 6 장기 기억과 웹 검색의 활용 (출처: BlenderBot)

필요가 있다. 이를 위해 수십 턴의 대화 문맥을 넘어 [그림 6]과 같이 오래전의 대화 정보까지 활용할 수 있는 장기기억 능력을 갖추어야 한다. 현재 장기기억과 관련된 연구는 현재 다양하게 진행되고 있는데 [15-17], 실제 챗봇 제품에 적용된다면, 챗봇은 대화 상대방을 더 잘 이해하고 친밀한 관계로 발전할 수 있을 것이다.

그리고 새로운 외부 지식이 빠르게 축적되고 있는 사회에서 사전 학습된 말뭉치에 없는 지식이 필요한 경우도 있다. 이러한 경우 실제 사람이 인터넷에 검색해서 지식을 습득하는 것과 같이 스스로 판단하여 웹 검색을 하여 외부 지식을 활용하여 대화에 이용할 수 있어야 한다. 실제 BlenderBot 등의 경우 [그림 6]과 같이 이러한 과정을 거쳐 외부 지식을 활용할 수 있는 기법을 제안하였는데 [8, 9], 앞으로의 지능형 챗봇의 경우 효과적이고 빠르게 웹 기반의 외부 지식을 활용하여 대화를 할 수 있어야 할 것이다.

웹 검색과 같이 챗봇이 외부 지식을 빠르게 습득하고, 학습 시의 데이터 분포와 실제 서비스상에서의 데이터 분포의 차이를 줄이는 방법으로 연속 학습(Continual learning)이 있다. 연속 학습이란 이전의 학습했던 지식을 잊지 않으면서 지속적으로 축적되는 데이터에 대해서 지식을 추가적이고 학습하는 방법론으로, 학습을 진행할 때 규제(Regularization)를 적용하거나 [20] 새로운 모델 파라미터를 추가하는 방식 [21, 22] 신규 학습 시 이전 학습 데이터를 같이 학습하는 방법 [23, 24] 증류 학습(Knowledge distillation)을 이용하여 신규 학습 시 이전 학습된 정보도 같이 전이하는 방법 [25] 등을 통해 이전 학습의 정보 잊지 않거나 학습된 파라미터 분포를 크게 변동하지 않도록 할 수 있다. 이러한 연속 학습을 이용하여 서비스를 통해 얻게 되는 사용자의 피드백 정보 등을 시의성 있게 반응할 수 있고, 새로운 지식을 학습하여 챗봇의 대화 성능을 지속적으로 개선할 수 있는 장점이 있다.

또한 사용자의 취향 및 선호에 따라 선택적으로 챗봇의 페르소나도 선택할 수 있어야 한다. 이를 위해 다양한 종류의 페르소나의 챗봇을 만들어야 하는데, 아직까지는 챗봇 하나를 개발하는데 시간 및 비용 등의 많은 자원을 필요로 한다. 따라서 새로운 페르소나의 챗봇을 개발함에 있어 효율적인 프로세스가 필요하고, 페르소나에 일관성 있게 대화를 할 수 있어야 한다. 최근 대형 생성 모델에 대해서 입력 프롬프트 (Prompt) 정보에 화자의 페르소나 정보를 넣어 가변적인 페르소나를 가능하게 하면서 일관성 있는 대화를 하는 연구들이 많이 진행되고 있다 [5, 6, 18].

그리고 현재의 챗봇은 주로 텍스트 정보만을 대화에 활용하고 있다. 이번 이루다 2.0에 탑재된 포토챗 기술처럼 이미지, 동영상 및 소리 등도 이해하기 위해서는 멀티 모달 (Multi-modality) 기술의 적용이 필수적이다. 특히 최근 Stable Diffusion [29]의 경우 주어진 텍스트에서 상당히 좋은 품질의 이미지를 생성할 수 있는데 [19], 사용자뿐만 아니라 직접 그림 혹은 사진을 교환하기 위해서는 이미지 생성 기술도 적용이 가능할 것이다. 또한 챗봇의 페르소나에 맞는 목소리를 탑재하거나 문자 메시지 입력 형식에서 벗어나 실제 통화까지 할 수 있는 챗봇을 구현하기 위해서는 음성 인식과 음성합성을 접목할 필요가 있을 것이다.

최근 초대형 생성 모델들이 기존 모델 대비 생성 품질이 비약적으로 발전하는 것을 보이면서 초대형 생성 모델을 서비스에 적용하려는 시도들이 많이 있다. 하지만 초대형 생성 모델은 파라미터의 개수가 매우 많기 때문에 추론하는 속도가 매우 느리고 그 비용이 큰 단점이 있다. 이를 극복하기 위해 증류 학습을 통해 초대형 생성 모델의 성능을 적은 모델에 전이하거나 [26], 파라미터를 가지치기 (Pruning) 하거나 [27] 양자화 (Quantization) 하여 [28] 모델의 성능 손실을 최소화하면서 그 크기를 줄이는 방식이 있다. 또한 마이크로소프트의 Deepspeed[6] 등의 프레임워크와 같이 연산 커널 퓨징 (Kernel fusing) 혼합 정밀도 변경, 다중 GPU 분산 추론 등 추론 최적화를 할 수도 있다. 추후 챗봇에 대해서도 초대형 생성 모델을 적용하고자 한다면 효율적인 추론을 위해 모델의 경량화 및 추론 최적화는 필수적일 것이다.

5. 결 론

본 원고에서는 일상 대화 챗봇에 대해서 전반적으로 다루었다. 1장에서는 여러 종류의 챗봇과 그 작동 원리를 살펴보았고, 2장에서 이루다 1.0에 비해 이루다 2.0이 개선된 점들을 살펴보았다. 이를 통해 일상 대화를 위한 오픈 도메인 챗봇이 사람들과 자유롭게 대화를 할 수 있을 정도로 수많은 발전을 거듭하였음을 알 수 있다. 물론 실제 사람과 같은 수준으로 고등한 대화 수준을 하기 위해서는 앞으로 해결해야 할 과제가 많이 남아있다. 이와 관련하여 4장에서 어떠한 문제가 있는지, 일상 대화 챗봇을 개선하기 위한 기술적인 방안들에 대해서 살펴보았다. 앞으로 일상 대화 챗봇은 단순히 시간을 소비하기 위한 오락적인 요소를 넘어서 사람들과 자유롭게 소통하고 서로의 일상 및 고민을 공유하고 위로를 받으면서 현대 사회에서 채울 수 없는 부족한 관계를 채워줄 있는 역할을 할 것이라 기대한다. 본 원고를 통해 일상 대화를 위한 오픈 도메인 챗봇을 연구하거나 적용하고자 하는 독자에게 도움이 될 것으로 기대한다.

6) https://www.deepspeed.ai/

참고문헌

[1] Weizenbaum, Joseph. "ELIZA—a computer program for the study of natural language communication between man and machine." *Communications of the ACM* 9.1: 36-45. 1966.

[2] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT*. 2019.

[3] Zhou, Li, et al. "The design and implementation of xiaoice, an empathetic social chatbot." Computational Linguistics 46.1: 53-93. 2020.

[4] Adiwardana, Daniel, et al. "Towards a human-like open-domain chatbot." arXiv:2001.09977. 2020.

[5] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33: 1877-1901. 2020.

[6] Thoppilan, Romal, et al. "Lamda: Language models for dialog applications." *arXiv:2201.08239*. 2022.

[7] Roller, Stephen, et al. "Recipes for Building an Open-Domain Chatbot." *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

[8] Shuster, Kurt, et al. "Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage." *arXiv:2208.03188*. 2022.

[9] Komeili, Mojtaba, Kurt Shuster, and Jason Weston.

"Internet-Augmented Dialogue Generation." *Proceedings of ACL*. 2022.

[10] Holtzman, Ari, et al. "The Curious Case of Neural Text Degeneration." *International Conference on Learning Representations*. 2019.

[11] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

[12] Xu, Jing, et al. "Recipes for safety in open-domain chatbots." *arXiv:2010.07079*. 2020.

[13] Dinan, Emily, et al. "Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack." *Proceedings of EMNLP-IJCNLP*. 2019.

[14] Li, Xiang Lisa, et al. "Contrastive Decoding: Open-ended Text Generation as Optimization." *arXiv:2210.15097*. 2022.

[15] Wu, Qingyang, et al. "Memformer: The memory-augmented transformer." *arXiv:2010.06891*. 2020.

[16] Martins, Pedro Henrique, Zita Marinho, and André FT Martins. "∞-former: Infinite Memory Transformer-former: Infinite Memory Transformer." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022.

[17] Xu, Jing, Arthur Szlam, and Jason Weston. "Beyond Goldfish Memory: Long-Term Open-Domain Conversation." *Proceedings ACL*. 2022.

[18] Madotto, Andrea, et al. "Few-Shot Bot: Prompt-Based Learning for Dialogue Systems." *arXiv:2110.08118*. 2021.

[19] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[20] Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences* 114.13: 3521-3526. 2017.

[21] Madotto, Andrea, et al. "Continual Learning in Task-Oriented Dialogue Systems." EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings. 2021.

[22] Qin, Yujia, et al. "ELLE: Efficient Lifelong Pre-training for Emerging Data." Findings of the Association for Computational Linguistics: ACL 2022. 2022.

[23] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." Proceedings CVPR. 2017.

[24] de Masson D'Autume, Cyprien, et al. "Episodic memory in lifelong language learning." NeuIPS. 2019.

[25] Jin, Xisen, et al. "Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora." Proceedings of BigScience Episode\# 5--Workshop on Challenges & Perspectives in Creating Large Language Models. 2022.

[26] Sun, Siqi, et al. "Patient Knowledge Distillation for BERT Model Compression." *Proceedings of EMNLP-IJCNLP*. 2019.

[27] Zaken, Elad Ben, Yoav Goldberg, and Shauli Ravfogel. "BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models." Proceedings of ACL (Volume 2: Short Papers). 2022.

[28] Whittaker, Edward WD, and Bhiksha Raj. "Quantization-based language model compression." *INTERSPEECH*. 2001.

[29] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of CVPR. 2022.

약 력

강 경 필

2016년 고려대학교 컴퓨터학과 졸업 (학사)
2020년 고려대학교 컴퓨터전파통신공학과 졸업 (박사)
2020년 ~ 스캐터랩 핑퐁팀 머신러닝 리서처
관심분야: 자연어처리, 데이터마이닝, 추천시스템
Email : rudvlf0413@gmail.com