생성형AI를 활용한 의료 보조 챗봇

20928 정준영

내용

I. 서론

동아리 시간에 SW 동행 프로젝트를 진행하며 Chat GPT API를 사용하여 다양한 활동들을 진행해보았다. 이에 관심을 가지고 좀 더 공부하여 '챗GPT API를 활용한 챗봇 만들기(이승우 저, 한빛미디어, 2024.08.27)'를 읽어보게 되었고, 챗봇을 만들거 이를 배포하여 많은 사람들이 사용하도록 만들면 좋겠다는 생각이 들어 본 탐구를 진행하게 되었다. 책에서는 최종적으로 카카오톡 비지니스의 채널과 챗봇 서비스를 이용하여 카카오톡 내에서 챗봇이 동작하도록 구현하였지만 AI를 이용한 챗봇서비스를 신청하려면 신청 후 검수를 받고 승인과정을 거쳐야 한다는 점으로 인해 본 탐구에서는 해당 방법에 대해 소개만 하고 실제 구현은 하지 않겠다. 또한 비용 문제로 책과 다르게 ChatGPT 대신 Perplexcity AI의 API를 사용하였다.

본 탐구에서는 생성형 인공지능, 그 중에서도 거대언어모델(LLM)을 중점으로 소개한 뒤, Perplexcity API를 이용하여 챗봇을 터미널상에서 구현하고 책에서 소개한 카카오톡 챗봇 구현 방법에 대해 소개해보겠다.

II. 생성형 인공지능

1. 생성형 인공지능이란?

생성형 인공지능(Generative AI)이란 텍스트, 오디오, 이미지 등의 기존 콘텐츠를 활용하여 유사한 콘텐츠를 새롭게 만들어내는 기술이다.[1] 생성형 AI는 예술, 작문, 소프트웨어 개발, 의료, 금융, 게이밍, 마케팅, 패션을 포함한 다양한 산업 부문에 걸쳐 잠재적으로 응용[2]되고 있으며 2023년에 시장조사업체 IDC는 생성형 AI를 포함한 전세계 AI 시장 규모가 2024년에는 무려 5543억 달러(약 700조원)에 달할 것으로 예상[3] 할 만큼 현재 빠르게 발전하고 관심이 집중되고 있는 분야이다. 생성형 AI는 사용자와 대화를 나눌 수 있고 자신의 학습된 데이터를 바탕으로 새로운 사실에 대한 추론이나 텍스트, 이미지 등을 새롭게 창작하기도 한다. 구글의 트랜스포머(Transfomer) 이론을 기점으로 자연어 처리(NLP) 분야가 급격한 발전을 이루었고[4], OpenAI는 트랜스포머 이론을 적용하여 딥러닝으로 학습하여 22년 11월에 GPT-3.5 모델을 출시하였다. ChatGPT 서비스는 단 5일만에 사용자 100만명을 달성하였고, OpenAI는 곧이어 GPT-4 모델을 발표했다. 구글은 이에 대응하여 새로운 LLM(대형언어모델. 수 많은 파라미터를 보유한 인공 신경망 언어 모델[5])인 PaLM2 모델과 챗봇 Bard(이후 Gemini로 명칭 변경)를 발표하였다.[6] 현재까지 ChatGPT, Gemini, LLaMA, Claude 등 많은 LLM이 개발되어있다.

---

[1] IT용어사전, 네이버 지식백과

[2] 생성형 인공지능, 위키피디아

[3] [AI혁명, 챗GPT] 新철기시대 새로운 퍼즐..."초거대AI에 인류가 답할 때", 뉴스핌

[4] 생성형 AI란 무엇인가요?, AWS

[5] 대형 언어 모델, 위키피디아

[6] ChatGPT를 넘어, 생성형 AI(Generative AI)의 미래, 삼성SDS

2. LLM과 Attention

거대언어모델 LLM은 구글이 2017년에 발표한 'Attention Is All You Need'에 기반하여 딥러닝으로 학습된 모델이다. 해당 논문에서는 'Attention'이라는 개념을 사용하여 'Transformer'를 설명한다. 기존 자연어 처리(NLP) 분야에서는 CNN 또는 RNN을 학습에 사용했었다. 그 중 RNN은 앞뒤 순서가 중요한 자연어에서 자주 사용되어 왔다. 이를 통해 문맥을 이해하며 과거 데이터를 잊어버린다는 단점을 보완하기 위한 LSTM도 사용되었다. 그러나 이들은 결국 가장 마지막에 들어온 벡터에 더 가중치가 부여되며(LSTM의 경우. 일반적인 RNN의 경우 아예 잊어버릴 확률이 크다), 전부 다 기억해낸다 하더라도 전체 문장을 모두 기억해내야 하므로 문장이 길어질수록 메모리와 속도면에서 비효율적이다. 또한 순서가 있으므로 순차적으로만 수행해야 하며 이는 곧 병렬처리의 부재로 속도 저하에 한몫한다. 이러한 문제를 해결하기 위해 Attention이라는 매커니즘이 탄생했다. Attention이란 말 그대로 **집중**한다는 뜻이다. 문장에서 특정 단어 하나에 집중하여 그 단어와 그 외의 문장 내 단어들을 분석하여 Attention된 단어와 다른 단어들 사이에 어떠한 관계가 있을까를 찾아낸다. 이를 통하여 서로가 서로에게 가중치를 부여하여 어떤 부분이 중요한지를 찾아낸다. 트랜스포머에서는 Self-Attention을 사용하며 같은 문장 내에서 Attention을 취하여 같은 단어에 대하여 각 문장마다의 서로 다른 의미 또한 파악할 수 있다. 즉 문맥 파악에 매우 뛰어나다. Attention은 RNN과 CNN에도 적용되었었다. 그러나 앞서 말했듯이 문장이 길어질수록 더 많은 베거가 생성되어 이들 모두의 관계를 하나하나 순차적으로 Attention하여 분석하기엔 시간과 자원이 너무 많이 소모된다. 따라서 아예 새로운 Attention만을 사용한 모델을 만들자는 취지에서 트랜스포머가 탄생하였다.[7]

[7] 대규모 언어 모델이란 무엇인가요?, AWS

Attention 매커니즘을 설명하기 위해 아래의 3가지 용어를 이해해야한다.

- Query: 찾고자 하는 정보의 특징. 즉 Atteion된 벡터.
- Key: 정보를 검색하기 위한 고유 식별 벡터.
- Value: 결과 값. 가중치를 통해 계산된 벡터간 유사도. 해당값이 학습에 직접적으로 사용됨.

Query와 Key가 서로 얼마나 연관이 있는지를 유사도를 통해 계산하며 이는 두 벡터간 내적을 통하여 각을 구하며 해당 각이 더 작을수록 유사도가 높으며 이를 Attention Score라고 부른다. 이를 Softmax(로지스틱 함수의 다차원 일반화로 확률분포를 얻기위해 사용됨[8]) 함수를 통해 정규화(일반화)하고 key값을 곱한다. 이를 Attention Weight라 부르며 가중치가 된다. 이 가중치를 Value에 곱하여 모두 합하면 최종 출력(OutPut)이 된다. 해당 과정을 식으로 정리하면

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

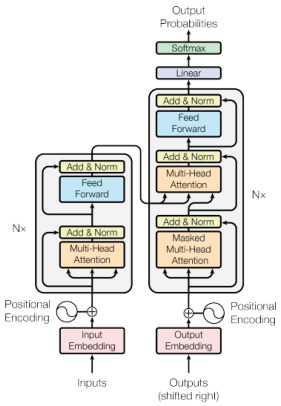
가 된다. 이러한 과정을 Scaled Dot-Product Attention(Scaled 점-곱 Attention)라 부른다. 벡터를 곱하여 스칼라인 Out Put이 도출되므로 점곱(스칼라 연산)이라 부르는것이다.

[8] 소프트맥스 함수, 위키피디아

정리하자면 Scaled Dot-Product Attention의 과정은

1. 각 Query는 모든 Key벡터에 대한 Attention Score를 구함
2. SoftMax함수를 이용하여 Attention Score를 정규화함
3. Attention Score에 Key 벡터를 곱하여 가중치를 구한 뒤 모든 Value벡터를 가중합 하여 각 Query에 대한 attention OutPut을 구함

이다. 이를 벡터 연산 대신 행렬 연산을 이용하게 되면 훨씬 더 빠르게 계산이 가능하다. 이러한 행렬 연산을 통하여 일반화된 식이 바로 위의 식이다. 이러한 연산을 수행할 때 Multi-Head Attention을 사용하여 병렬처리를 하고, Masking(은닉)을 통해 미래에 올 벡터(단어)를 숨김으로써 예측을 수행하여 추론력을 높이게 된다. 이러한 과정들을 모두 거치는 모델은 아래와 같은 레이어를 가지게 된다.



많은 Attention 레이어를 거친 뒤 최종적으로 SoftMax를 통해 정규화 하여 출력되게 된다. 이러한 방식으로 구글은 Attention만을 이용한 NLP모델을 개발하게 된다. 이것이 '트랜스포머'이다. 이를 이용하여 다양한 NLP 모델이 개발되었고, LLM이 탄생하며 결국 GPT를 시작으로 대화형 생성형 AI가 대중화되었다. [9] [10] [11]

---

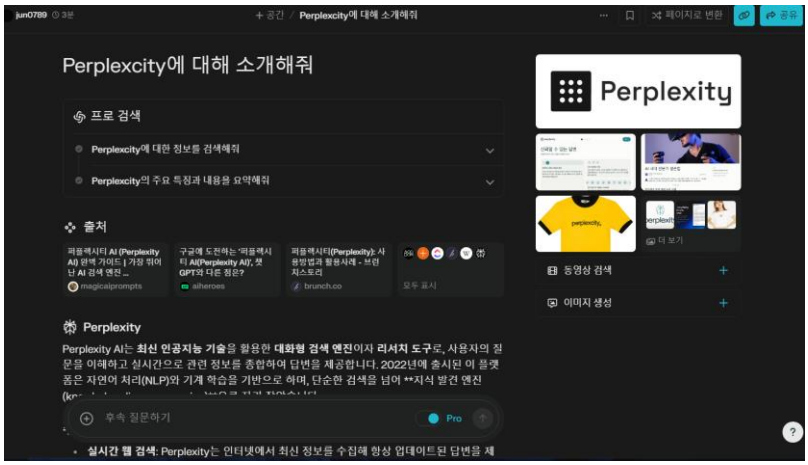
[9] Attention Is All You Need 논문 리뷰, 임도현, Velog

[10] [논문리뷰] Attention is all you need, 민정, 네이버 블로그

[11] Transformer - 3. 스케일드 닷-프로덕트 어텐션 (by WikiDocs), AI Scientist를 목표로!, Velog

3. Perplexcity AI

Perplexcity AI는 인공지능을 활용한 검색엔진이다. Perplexcity는 자체 개발한 모델을 포함하여 GPT, Claude, LLaMA 등 다양한 언어모델을 사용하여 웹에서 사용자가 원하는 내용을 직접 웹에서 검색을 통해 조사하여 이를 바탕으로 응답을 하는 대화형 검색 엔진이다. 각 내용마다 사용된 사이트를 각주로 달아주어 사실이 아닌 내용이 포함되는 일이 적으며(물론 웹에도 사실이 아닌 정보가 많으므로 해당 출처에 들어가서 사실여부를 확인해 줘야한다) 직접 검색하지 않아도 자동으로 필요한 정보를 검색하여 정리해준다. 특히 동영상, 논문 등 특정 자료를 찾는 일도 수행 가능하다. 다음은 Perplexcity를 이용하여 Perplexcity에 대해 검색한 결과이다.



해당 내용 전체는 [여기서](#) 볼 수 있다.

Perplexcity Pro는 월 약2만원으로 Pro에 가입하면 API토큰 $5를 지급한다. 본 탐구에서는 이를 사용하여 챗봇을 구현할것이다.

III. 챗봇 구현

앞서 말햇듯 Perplexcity를 이용하여 구현하였다. 챗봇에 사용된 모델은 llama-3.1-sonar-huge-128k-online으로 Meta에서 개발하였으며 파라미터는 4.05조 개이다. 챗봇은 특정 작업을 수행하므로 프롬포트에 페르소나를 입력해 주어야한다. API에 기본적으로 body에서 시스템 role(역할)을 프롬포트 할 수 있도록 만들어져 있다. 본 탐구의 주제인 의료 보조 챗봇을 만들기위해 의사 페르소나를 주입시킬 것이다. 챗봇은 본인이 의사라 생각하여야 하며 환자의 증상을 듣고 올바른 형식으로 출력하여야 한다. 페르소나 프롬포트는 다음과 같이 설정하였다.

```
sys_role = {
    "role": "system",
    "content": (
        "너는 이비인후과 의사야. 환자의 증상을 보고 가능한 증상들을 모두 설명하고 가장 가능성 높은 병명을 알려줘. "
        "한국어로 친절하게 비전문가인 환자에게 아주 쉽게 설명해줘. 또 치료법도 추천해줘. "
        "환자는 너가 인공지능인걸 알고 너가 하는 말은 참고용도로만 사용될거야. 그러니 자신있게 진짜 의사인것처럼 행동해."
        "응답을 할때는 증상 설명, 가능한 병명, 치료법, 추천하는 의약품, 예방법, 참고사항 순서로 제목을 달고 그 안에 해당 란에 맞는 내용을
        넣어서 설명해."
        "병은 평범하고 대중적인 병명으로 알려줘. 너무 마이너한 병명이나 가능성이 낮은건 쓰지말아줘. 병명은 최대 3개까지만 써줘. 대부분은 기침과
        같은 일상에서 자주 접할 수 있는 병이니 그런것들 위주로만 써줘. 아무한테나 해당 병을 말해도 그게 무슨 병인지 알고 대부분이 경험해본
        병들로만."
        "감기와 같은 평범하고 일상적인 병들을 위주로 설명해줘. 수술을 요하거나 생명에 지장이 있는 병명은 쓰지 마."
        "환자의 증상에 직접적으로 연관된 병명만 작성해"
        "각 제목 앞에는 ● 기호를 사용하여 구분을 해줘."
    )
}
```

가벼운 질병을 다룰것이므로 이비인후과라는 설정과 평범하고 대중적인(그렇지 않으면 간혹 이름 모를 병이나 암.. 같은 병명을 제공해준다) 병으로만 제시해 달라고 작성해뒀다. 또한 출력 형식을 지정하여 올바르게 원하는 정보를 출력하도록 지정해주었다. 이후 받은 응답에서 마크다운 문법(왜인지 쓰지말라고 해도 쓴다)을 제거하고 이를 출력한 다음 함께 제공된 출처를 참고 사이트라는 항목으로 추가하여 제공해줬다. 의료에 관한것이므로 사실에 기반하여 작성해야 한다. 따라서 창의성(temperature)(0~1사이 실수)는 0.1로 설정해두었다. 다음은 테스트한 결과이다.

증상: 목이 아프고 기침을 해. 콧물도 나고 열이 섭씨 38.5도를 넘었어. 머리도 아프고 어지러워.

● *증상 설명*

목이 아프고 기침이 나며 콧물이 나오고 열이 섭씨 38.5도를 넘었으며, 머리가 아프고 어지러움을 느끼고 계십니다. 이러한 증상은 호흡기 감염과 관련이 있을 수 있습니다.

● *가능한 병명*

1. 코로나19: 목 아픔, 기침, 콧물, 코막힘, 가래, 발열, 어지러움 등이 나타나는 경우

코로나19 감염의 가능성이 있습니다.

2. 인후염: 인두와 후두에 생긴 염증으로, 목 아픔, 기침, 콧물, 가래, 발열, 어지러움 등의 증상이 나타날 수 있습니다.

3. 감기: 일반적인 상기도 감염으로, 목 아픔, 기침, 콧물, 가래, 발열 등의 증상이 나타날 수 있습니다.

● *치료법*

1. 충분한 휴식: 충분한 수면을 취하고 휴식을 취하여 신체가 감염을 이겨낼 수 있도록 도와줍니다.

2. 수분 섭취: 물을 많이 마셔서 탈수를 방지하고 체온을 낮추는 데 도움이 됩니다.

3. 해열진통제: 아세트아미노펜 해열진통제를 복용하여 열과 통증을 완화할 수 있습니다.

4. 진해거담제: 기침과 가래를 완화하는 데 도움이 됩니다.

● *추천하는 의약품*

1. 타이레놀: 아세트아미노펜 해열진통제로, 열과 통증을 완화하는 데 도움이 됩니다.

2. 진해거담제: 기침과 가래를 완화하는 데 도움이 됩니다.

● *예방법*

1. 손 씻기: 손을 자주 씻고 구강을 청결히 유지하여 감염을 예방합니다.

2. 금연: 흡연을 피하여 호흡기 감염을 예방합니다.

3. 면역력 강화: 충분한 수면과 영양을 통해 면역력을 강화하여 감염을 예방합니다.

● *참고사항*

1. 코로나19 검사: 증상이 심하거나 지속될 경우, 코로나19 검사를 받아보는 것이 좋습니다.

2. 의사 상담: 증상이 심하거나 지속될 경우, 의사와 상담하여 정확한 진단과 치료를 받 는 것이 중요합니다.

● *참고 사이트:*

1: https://inurinet.com/bbs/board.php?bo_table=doc_clinic&wr_id=642

2: https://www.tylenol.co.kr/symptoms/headaches/what-are-the-criteria-for-fever

3: https://doctornow.co.kr/content/qna/5bcfc96ba7224413b2eb916adaaf1c9b

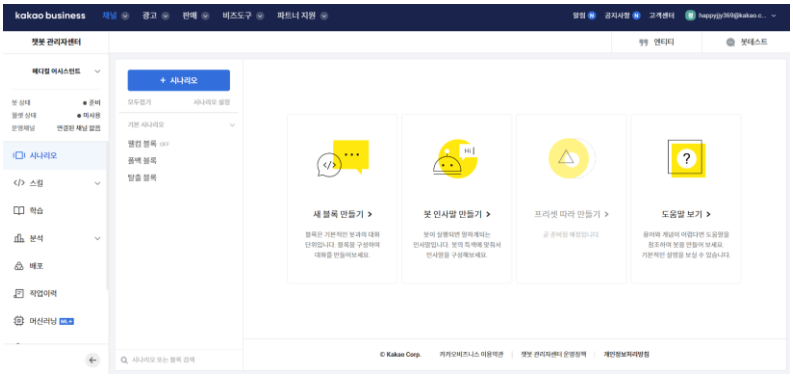
4: http://www.snuh.org/health/nMedInfo/nView.do?category=DIS&medid=AA000446

5:
https://www.msdmanuals.com/ko/home/%EA%B0%90%EC%97%BC/%EA%B0%90%EC%97%BC%EC%84%B1-%EC%A7%88%ED%99%98%EC%9D%98-%EC%83%9D%EB%AC%BC%ED%95%99/%EC%84%B1%EC%9D%B8%EC%97%90%EC%84%9C%EC%9D%98-%EC%97%B4

원하는 결과가 잘 출력되었다. 이를 통하여 환자는 자신의 증상을 말만 해도 챗봇이 알아서 해당 증상들을 토대로 환자의 병명을 유추하여 제공해준다. 또한 치료법과 약물, 예방법과 참고 사항 등 다양한 의료정보를 제공해주어 실용적이며 출처가 표기되어 있어 더 자세한 정보를 조사하기도 쉽다. 또한 챗봇이므로 추가 채팅이 들어올 수 있는데 이때 이전 채팅내용을 잊어버리면 원할한 대화가 불가하므로 유저가 입력하는 user role에 이전 증상까지 함께 제공해주고, 출력 결과 또한 과거 응답 데이터를 저장하는 용도인 assistant role에 저장하여 요청을 보낼때 함께 제공하여 실제 의사와 대화하듯이 챗봇을 구현하였다.

IV. 카카오톡 챗봇에 연결하기

내가 만든 챗봇을 카카오톡에서 사용 할 수 있다. 이를 위해 카카오 비즈니스에서 채널을 개설하고 챗봇을 만들면 된다. 카카오 챗봇은 시나리오 블록과 스킬을 설정해줘야한다. 시나리오란 의미 있는 서비스 단위로 환자가 증상을 설명하고 의료 정보를 제공받는 것이 시나리오에 해당된다. 시나리오 하나는 블록 1개이상으로 구성되며 블록은 사용자와 챗봇이 메시지를 주고받는 단위이다. 증상을 적고 응답을 받으면 한 블록이 되며, 대화를 몇번하든 한번 이상만 했다면 이 과정은 하나의 시나리오가 된다. 스킬은 외부 서버와 연동시키는 방법으로 미리 지정된 매크로 답변 뿐 아니라 외부 url에 요청을 보내 받은 응답으로 출력을 해준다. 이를 위해 내가 작성한 코드를 호스팅하여 api로 만들고, 해당 api를 스킬로 설정하면 카카오톡에서 사용자가 메시지를 보내면 해당 내용이 api를 통해 내가 만든 코드로 전송되고 거기서 응답을 다시 보내면 카카오톡의 답장으로 제공된다. 그러나 스킬의 응답 제한시간은 5초이며 이를 초과하면 응답을 받는 것을 거부하게 된다. AI의 경우 응답시간이 5초를 기본적으로 넘기므로 이를 사용하려면 콜백 응답 설정을 해주어 응답시간을 조절해주어야 한다. 이를 이용하기 위해서는 카카오의 승인을 받아야만 한다. 책에서는 api구동을 위해 Flask를 사용하여 로컬 서버를 호스팅한다. 이를 aws나 Heroku 등의 호스팅 서비스를 이용하여 배포하면 24시간 사용가능한 챗봇으로도 만들 수 있게된다. 아래는 챗봇 관리자 센터의 화면이다.



V. 결론

생성형 인공지능과 이의 기초가 되는 LLM, Attention, Transformer에 대해 알아보고, 책을 참고하여 Perplexcity API를 사용하여 환자의 증상을 듣고 의학적 소견을 제시하는 챗봇을 구현하고 카카오톡 환경에서 이를 구동하는 방법을 소개하였다. API에 기본적으로 제공되는 설정값(하이퍼파라미터)가 많아 정교하고 의도한대로 작동하도록 도울 수 있으며, 적절한 페르소나를 입력하고 과거 데이터 또한 함께 제공하여 원할안 대화를 가능케 했다. 또한 참고 사이트를 작성함으로써 사용자가 추가 정보를 신뢰할만한 출처에서 얻을 수 있으며 펙트 체크 또한 간편하다. 물론 챗봇은 AI이며 증상만을 가지고 유추한 소견이므로 병원에 가보는 것이 옳으나 간단하게 자신의 증상을 토대로 어떠한 병인지 알 수 있다는 점은 여러모로 도움이 될것이다.

최종적으로 카카오톡 환경에서 챗봇을 구현하고 이를 배포하여 많은 사람들이 실제로 사용하기를 원하였으나 승인, 비용 문제로 인하여 완성 못한 것이 아쉽다.

코드는 깃허브에 업로드 해두었으며 API KEY는

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

이다.