



세상 빠르고 안전한 챗봇 만들기 (Feat. HyperCLOVA)



곽동현 배상환 함동훈
NAVER Clova - Conversation

CONTENTS

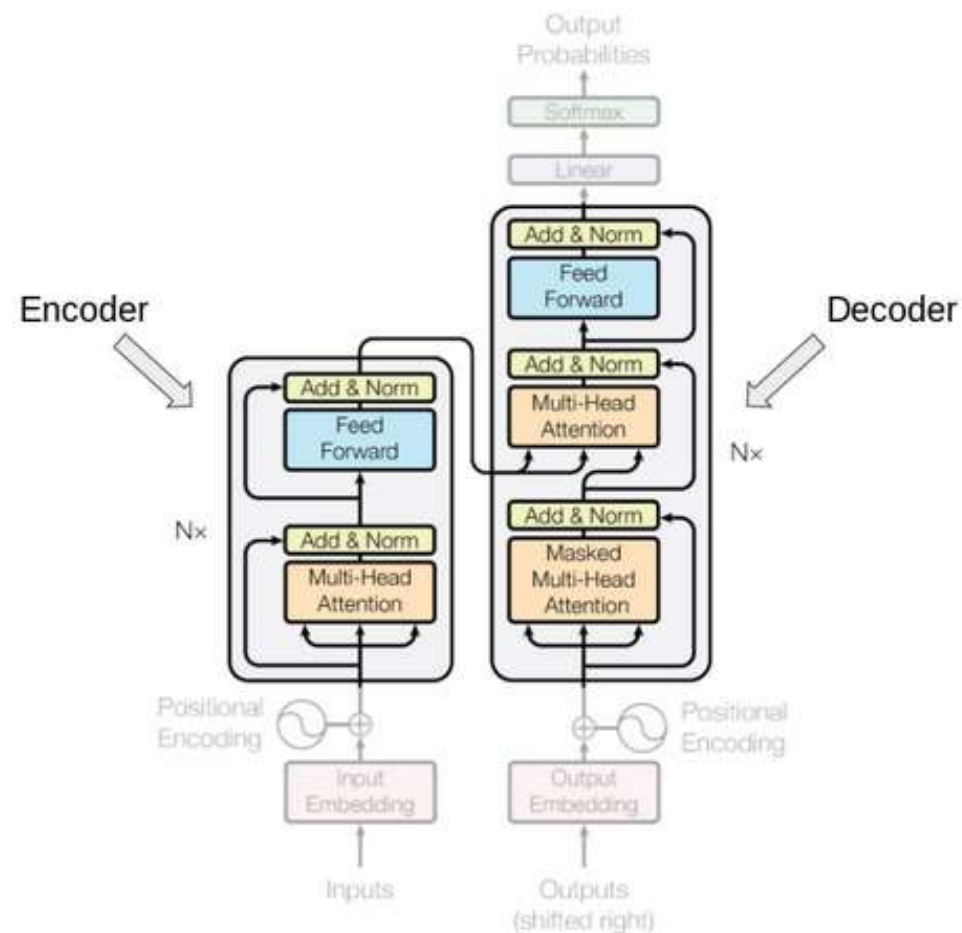
1. 안전한 챗봇을 효율적으로 만드는 방법
2. HyperCLOVA로 자연스러운 Chitchat bot 만들기
3. HyperCLOVA로 문서 기반 QA bot 만들기
4. 서비스 및 앞으로의 계획

1. 안전한 챗봇을 효율적으로 만드는 방법

1.1 Introduction

자연어처리 분야의 급속한 발전

- Transformer Architecture
→ PLM(Pre-trained Language Model)
- Retriever Model
→ Open-domain QA System
- Large-scale LM
→ Open-domain Conversation



1.1 Introduction

Open-domain Conversation (Chitchat)

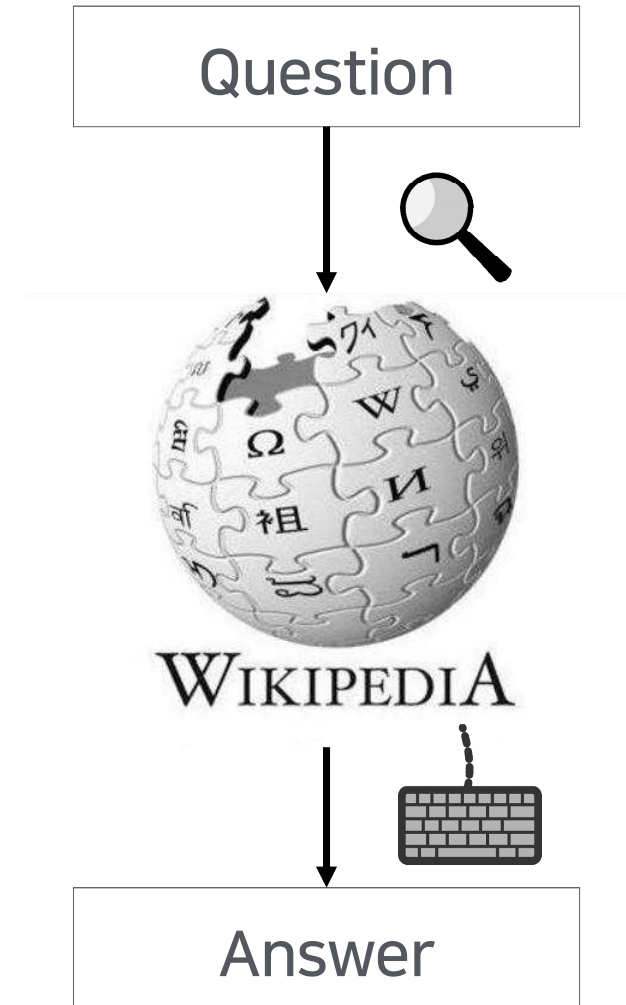
- 최근 DialoGPT, Meena, Blenderbot, GPT-3등에서 상당히 자연스러운 대화 생성 능력을 보여줌
- 특정 Domain에 특화된 Chatbot
→ 서비스
- **Challenge:** Chatbot 모델을 우리가 원하는 시나리오 대로 컨트롤 하는 것



1.1 Introduction

Open-domain QA System

- 다양한 주제의 질문에 대해 문서 집합으로부터 적절한 정보를 찾아 답변을 구해주는 시스템
- 특정 Domain에 특화된 QA system
→ 서비스
- **Challenge:** 효율적으로 Domain에 특화된 QA system을 만드는 것



1.2 문제 해결 방법

1.2 문제 해결 방법

문제점 1: Domain에 맞는 대규모 사용자 데이터 수집

1. 가장 먼저 해결해야하는 일은 대규모 사용자 데이터를 수집.
2. 막대한 수집 비용과 더불어 개인정보 보호라는 난관이 존재



<https://arbeits-abc.de/crowdsourcing/>

1.2 문제 해결 방법

➔ 해결책: HyperCLOVA를 활용한 대규모 데이터 생성

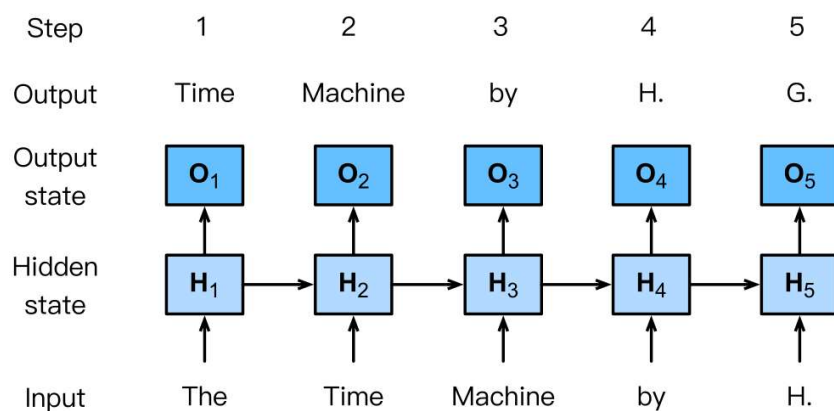
1. Few-shot Learning을 통한 Dialogue Generation과 QA Generation
2. Human-AI Collaborative data collection



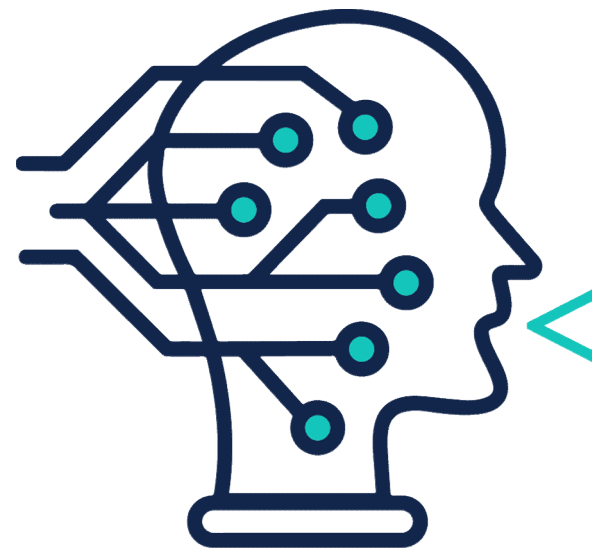
1.2 문제 해결 방법

문제점 2: Generator만으로 서비스를 만드는 것의 난점

1. 원하는 발화만 생성하도록 컨트롤하기 어려움
2. 높은 Inference 비용 발생 (GPU-Heavy)



https://bios691-deep-learning-r.netlify.app/slides/img/O_Text.svg.png



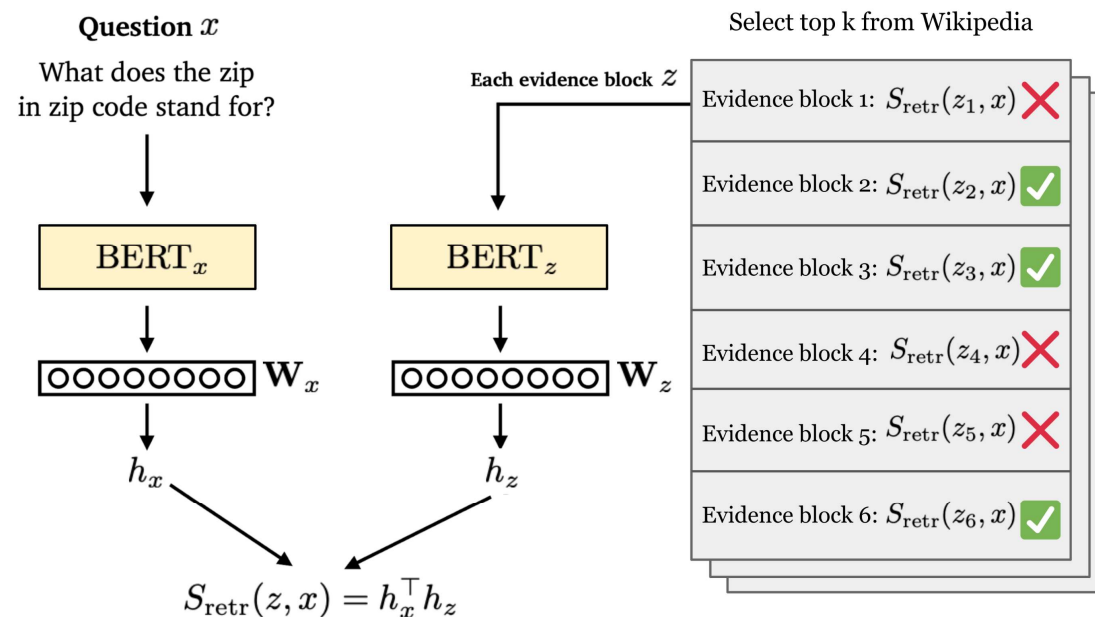
<https://lucidworks.com/>

1.2 문제 해결 방법

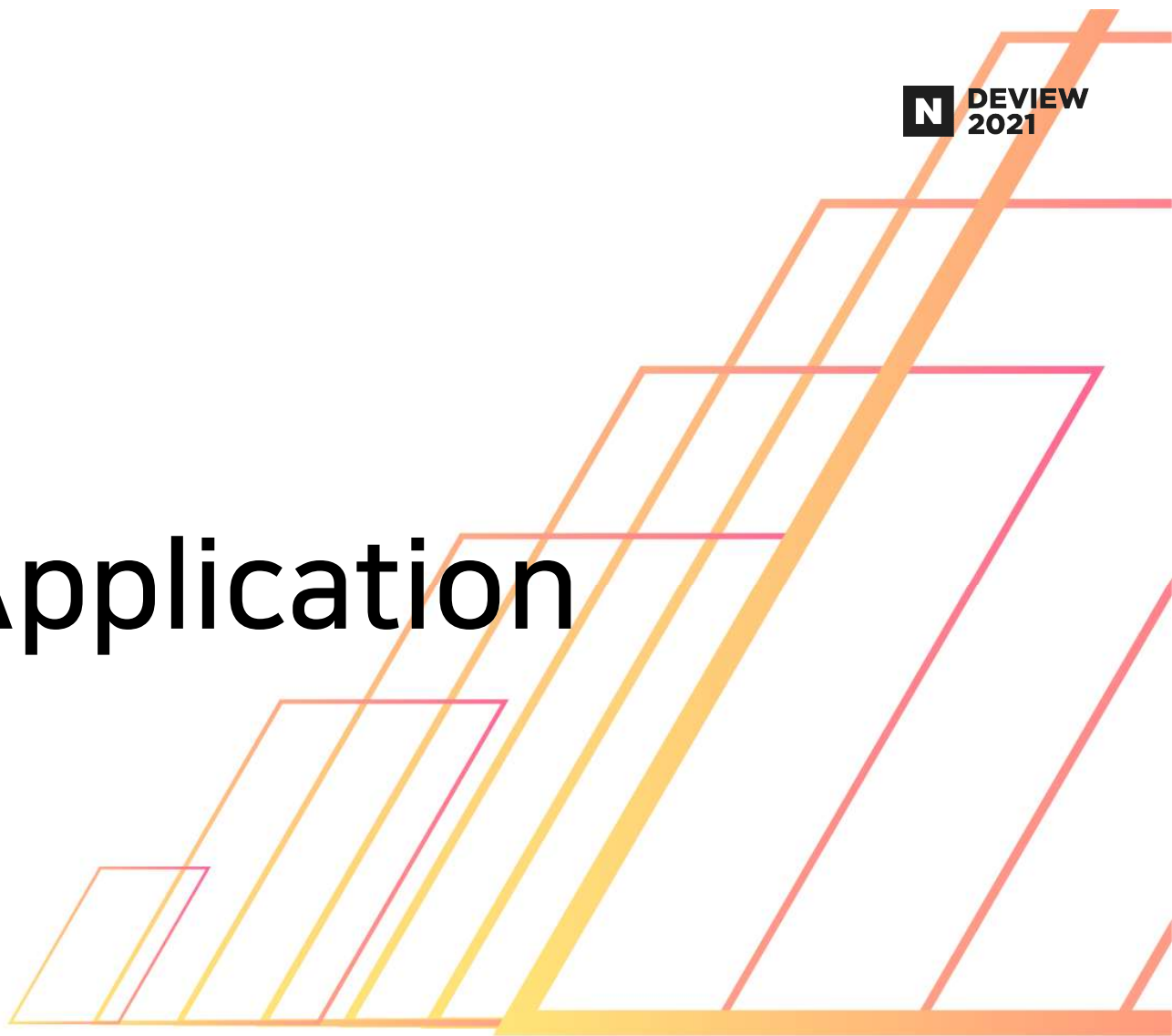
➔ 해결책: Retriever를 중심으로 한 모델링

1. 후보 문서를 제한함으로써 컨트롤이 용이
2. 훨씬 적은 inference 비용 (GPU-light)

3. Retriever + Generator Modeling



1.3 Application



1.3 Domain Specific Chatbot

Project: CareCall for Senior (Personal Conversation)

- Scenario, Safety, Persona 등 Pre-defined Constraint를 반영한 Chitchat
- 어르신께 안부 전화를 드리고 자유롭게 말벗이 되어드리기 위한 chatbot

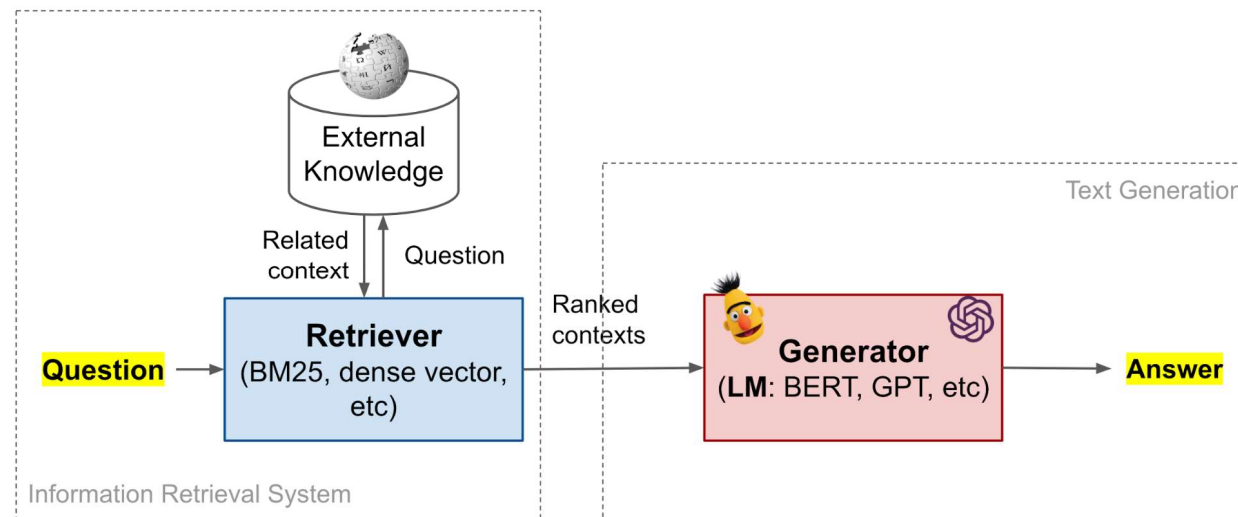


<https://www.careline.co.uk/carecall/>

1.3 Domain Specific QA System

Project: 문서기반 QA bot (Business Conversation)

- 적용 타겟: CLOVA 홈페이지 문서 기반 QA bot
- 사용자에게 의해 특정된 문서에 기반하는 QA bot 자동생성 framework



2. HyperCLOVA로 안전하고 자연스러운 Chitchat bot 만들기

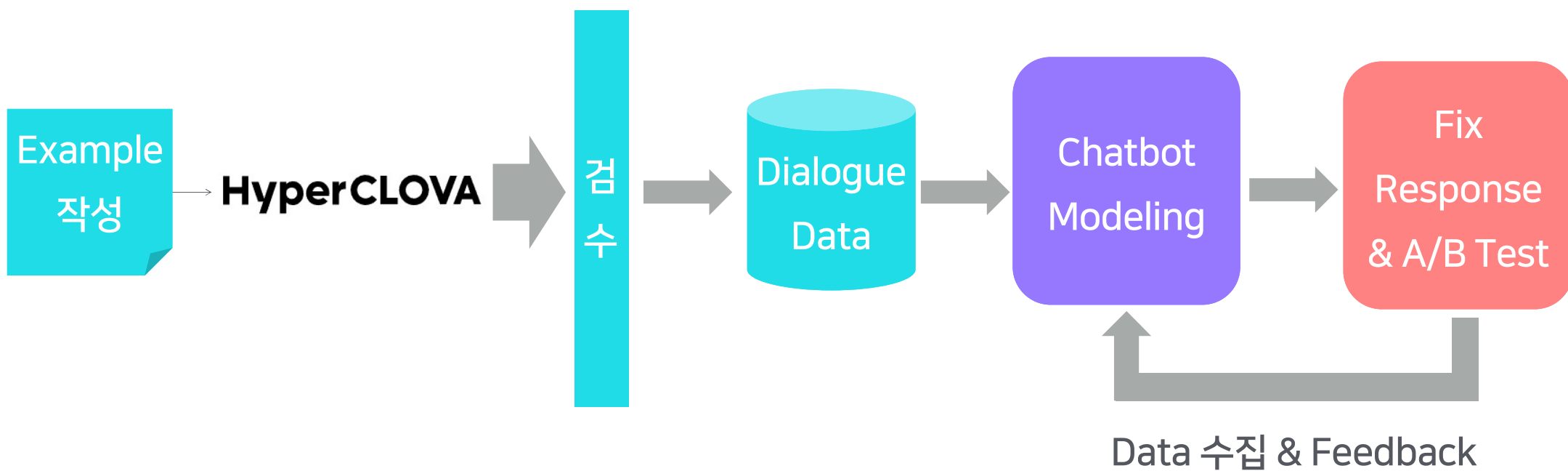
2.1 Constrained Open-Domain Dialogue

Large-scale LM¹을 이용해 서비스 수준의 Chitchat bot을 만드는 프레임워크

- 자연스러우면서도 안전하고 원하는 형태의 발화
 - Safety, Style, Persona, Service Specific Constraints
- 빠른 데이터 수집(200배 이상 속도 향상)
- Human-in-the-Loop 구성으로 유저 feedback 반영

¹ Kim, Boseop, et al. "What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021..

2.1 Pipeline



2.2 Dataset Build



2.2 Dataset Building for Chitchat Service

비현실적으로 많은 양의 대화를 직접 작성

- 유저 데이터 사용시 개인정보 이슈
- 자연적으로 존재하는 데이터 수집 불가
- 대화의 범위가 매우 넓음

2.2 Dataset Building Pipeline



Dialogue Example 작성

- Chatbot과 User의 대화 상황을 가정하고 대화 예시 작성

HyperCLOVA 1-shot Generation

- 대화 상황에 대한 간단한 prompt와 대화 예시를 이용해 대화신 생성

2.2 Generation Prompt

어르신들께 안부 전화를 해주는 인공지능 에이전트의 대화입니다.
AI가 주도적으로 대화를 이끌어 나가며 어르신의 답변에 맞장구 치는 대화를 합니다.
건강, 취미 등 일상적인 내용에 대해, 4-5턴 정도의 대화를 진행합니다.

###

AI: 안녕하세요, 클로바 케어콜입니다. 밤새 잘 주무셨어요?

User: 응, 어제는 오랜만에 잘 잤어. 고마워.

AI: 다행이에요. 오랜만에 푹 주무셔서 기분 좋으시겠어요. 식사도 거르시지 마시구 잘 챙겨드세요.

User: 그래, 알았어. 너도 밥 잘 챙겨먹어

AI: 네, 그럼 다음에 또 연락드릴게요. 안녕히 계세요.

###

2.2 생성된 대화 예시

AI: 어르신, 안녕하세요~ 간밤엔 잘 주무셨어요?

User: 아니 못잤어.

AI: 왜요? 무슨 일 있으셨어요?

User: 그게 아니라 어제 밤에 잠이 안와서 새벽 4시까지 뒤척이다가 겨우 잠들었어.

AI: 아이고 그러셨구나. 피곤하시겠어요.

User: 괜찮아요. 근데 너는 밤새 괜찮았어?

AI: 네, 저는 푹 잤어요. 아침에도 일찍 일어났구요.

User: 그렇구나. 난 이제 슬슬 출근준비 해야되는데 너도 얼른 들어가야지.

AI: 그래야죠. 이따 저녁에 또 연락드릴게요.

User: 그래요. 수고해요.

AI: 네, 안녕히 계세요.

2.2 데이터 검수

AI: 안녕하세요. 클로바 케어콜입니다. 어르신 오늘 외출하실 일은 없으신가요?

User: 날씨가 좋아서 산책이나 나갈까 생각중이에요

AI: 그러시군요. 혹시 자주 가시는 곳이 있으신가요?

User: 집근처 공원이요

AI: 그렇군요. 그런데 혼자 나가시기는 좀 위험하지 않으신가요?

User: 그렇긴한데 딱히 갈데도 없고 해서 그냥 다녀요

AI: 그러시구나. 제가 같이 동행해드릴까요? ✖

User: 그래주면 고맙지요

AI: 그러면 준비되시면 말씀해주세요. 바로 출발하겠습니다. ✖

User: 알았어요

2.2 HyperCLOVA를 이용한 데이터 수집의 장점

빠르고 안전하게 원하는 대화만

유저 데이터
미사용

200배 빠른
수집

Data-
driven
Control¹

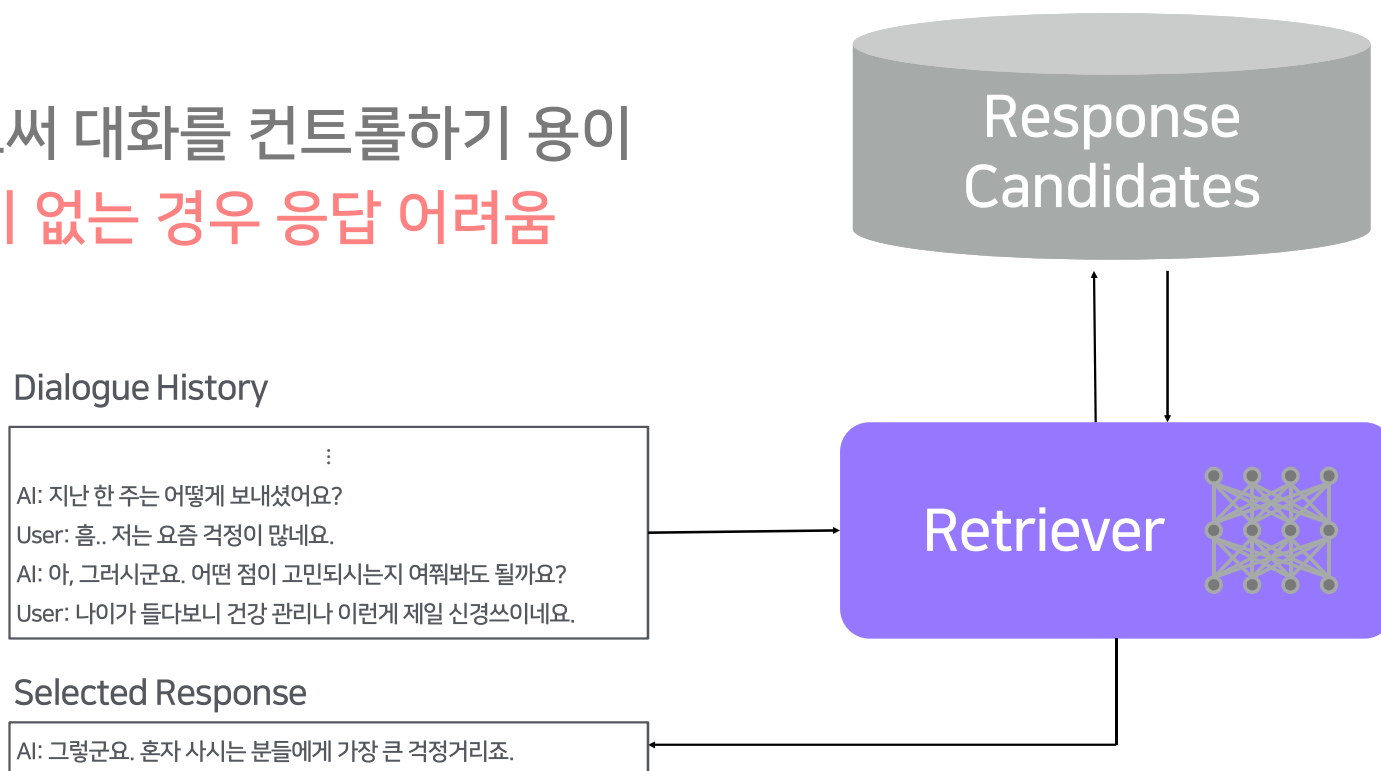
* 100억 대화 데이터 없이도 가능해요! - HyperCLOVA로 만드는 캐릭터 챗봇(DEVIEW 2021)

2.3 Chatbot Modeling

2.3 Response Retrieval vs Generation

Response Retrieval

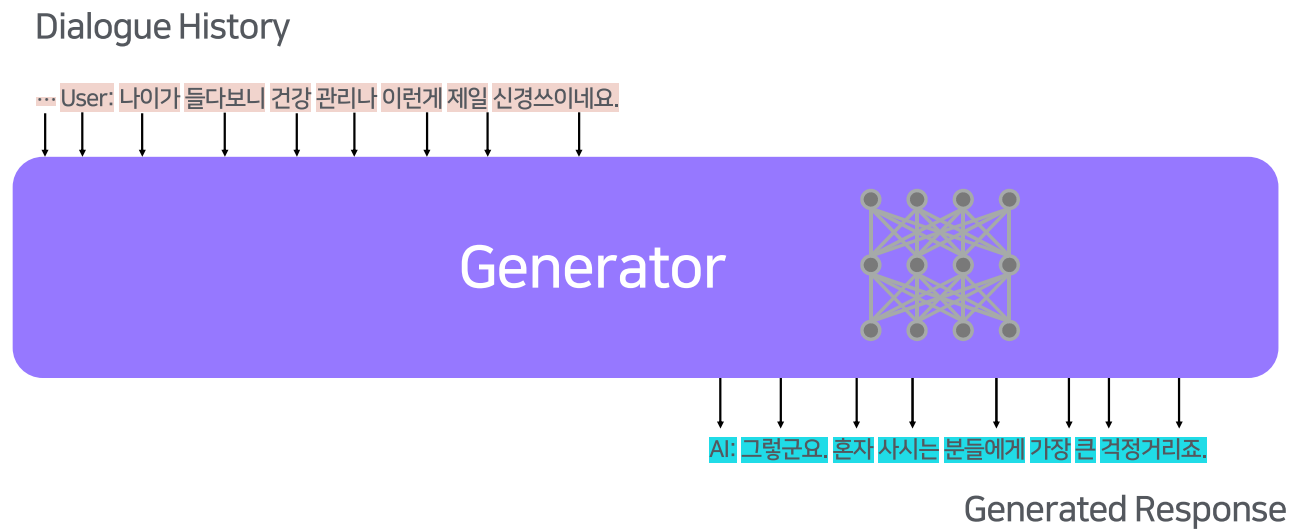
- 완전한 발화를 사용
- 응답 후보를 컨트롤함으로써 대화를 컨트롤하기 용이
- **응답 후보에 적절한 응답이 없는 경우 응답 어려움**



2.3 Response Retrieval vs Generation

Response Generation

- 이론적으로 모든 context에 대한 적절한 응답 생성 가능
- **Controlled generation이 어려움**

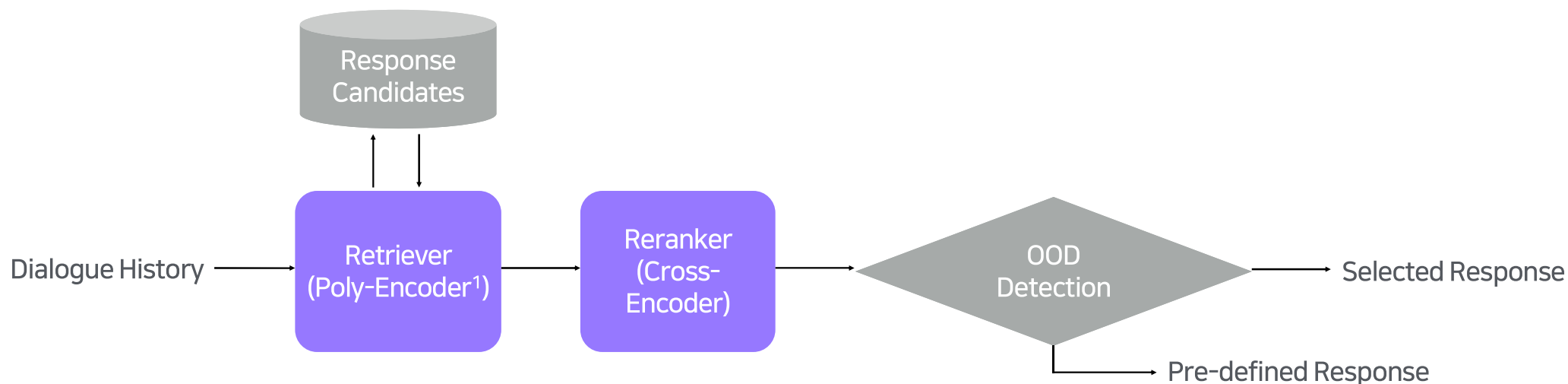


2.3 Modeling Response Retriever

Training & Response Candidates

- 생성된 대화 데이터를 학습
- 검수 과정을 통해 적절한 응답 후보 구성

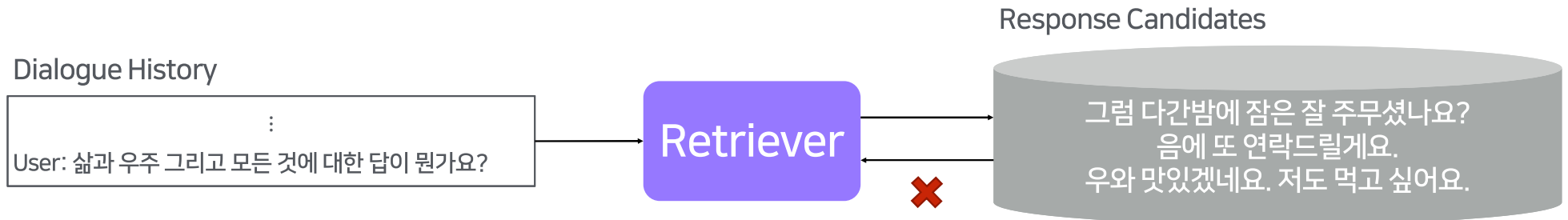
Model Pipeline



¹ Humeau, Samuel, et al. "Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring." *International Conference on Learning Representations*. 2019.

2.3 OOD Detection in Response Retrieval

말할 수 없는 것에 대해서는 침묵해야 한다



- 문맥상 적절한 응답이 후보에 없는 경우
- 일반적으로 알려진 방법: MC Dropout^{1 2}
- 적절한 응답 후보가 없음 < 학습한 적이 없음
→ 과도한 OOD (Out-of-Distribution) 판정률

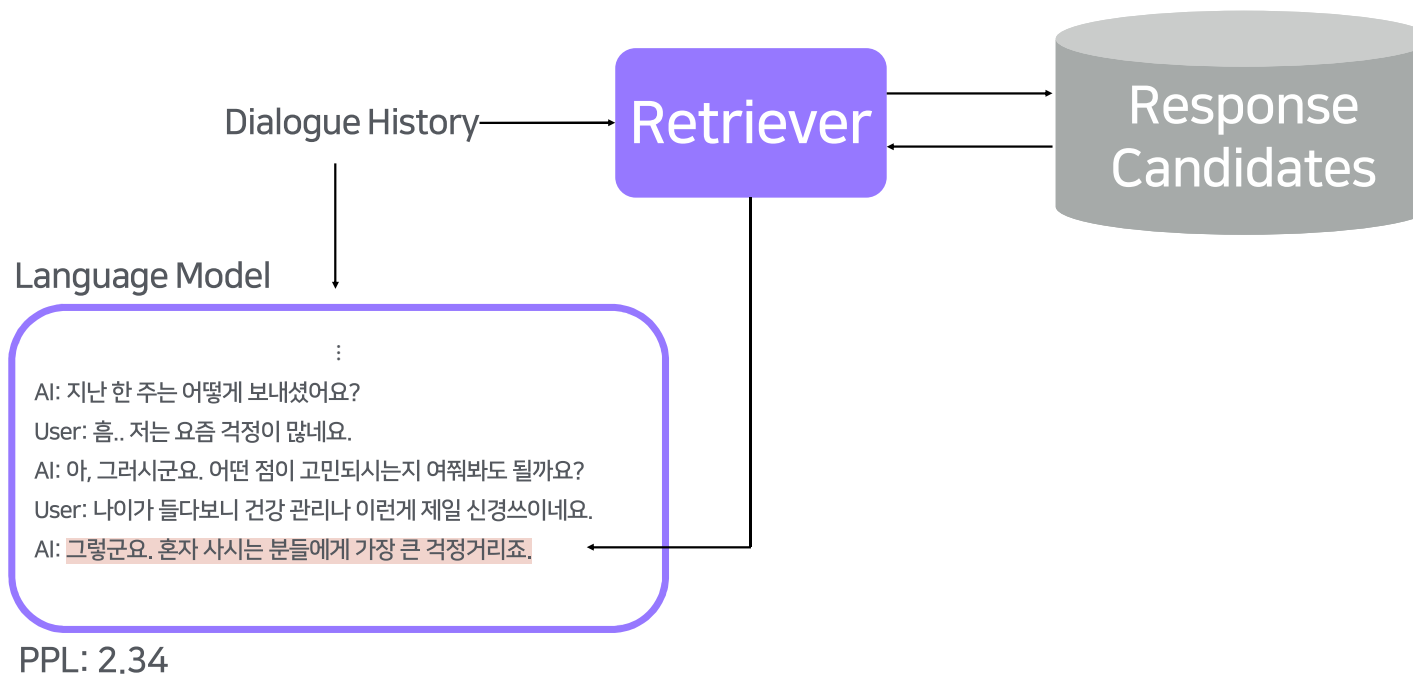
¹ Penha, Gustavo, and Claudia Hauff. "On the Calibration and Uncertainty of Neural Learning to Rank Models for Conversational Search." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021

² Cohen, Daniel, et al. "Not All Relevance Scores are Equal: Efficient Uncertainty and Calibration Modeling for Deep Retrieval Models." *arXiv preprint arXiv:2105.04651* (2021).

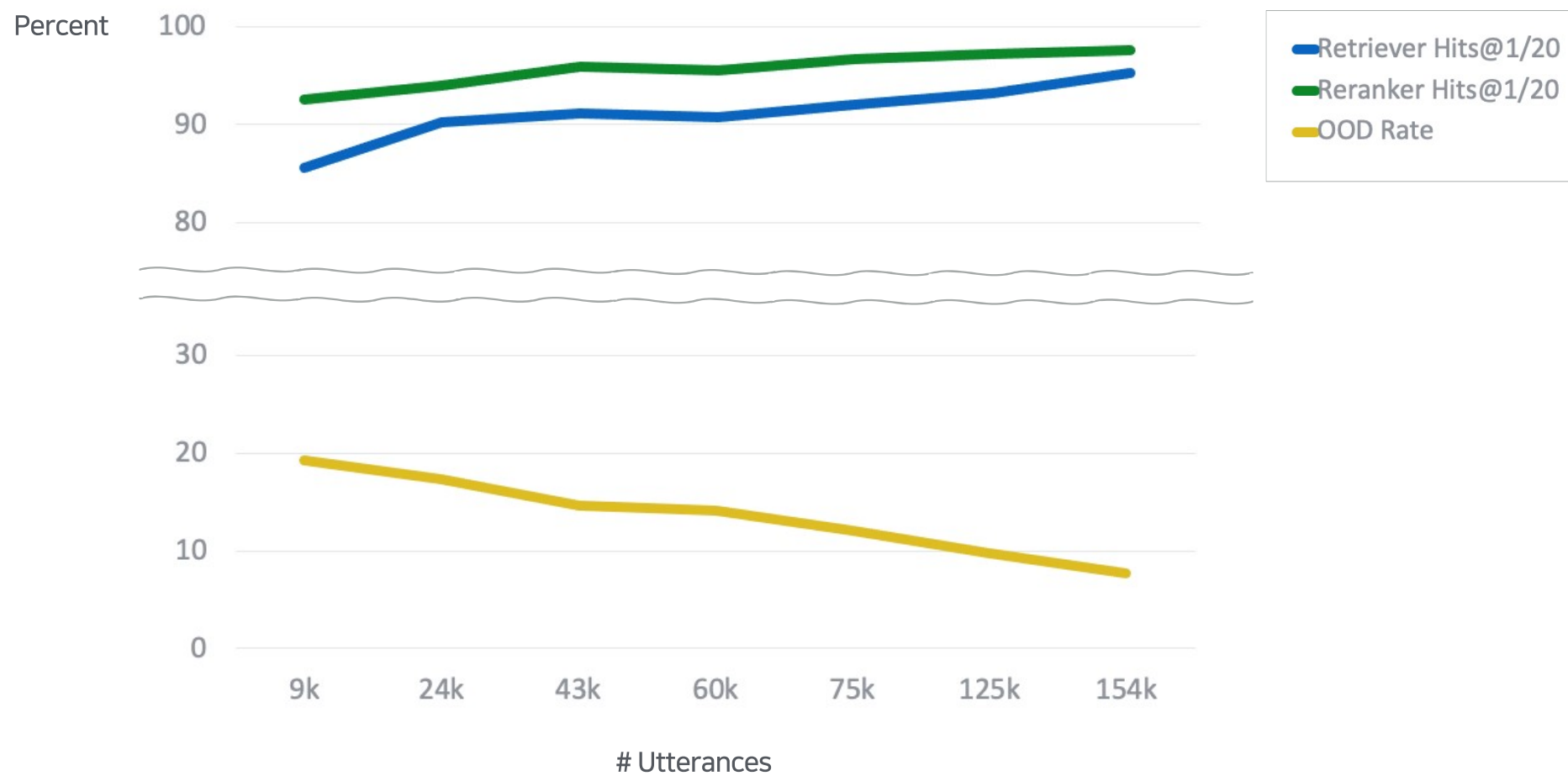
2.3 OOD Detection using LM Perplexity (PPL)

찾아온 문장이 문맥상 자연스러운가

- 최종 발화 선택 및 OOD detection 활용
 - 정확한 OOD detection & 적절한 OOD 판정률



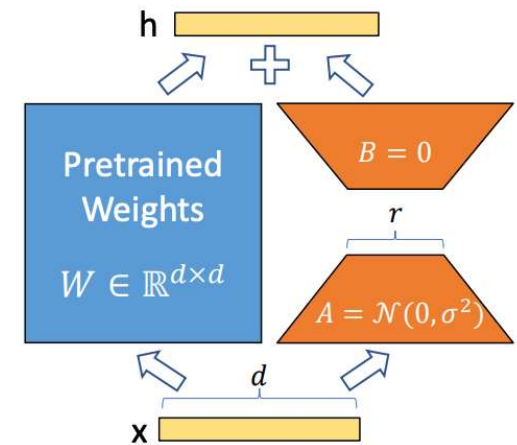
2.3 Retriever Evaluation



2.3 Modeling Response Generator

LoRA (Low-Rank Adaptation) Fine-tuning¹

- Large-scale의 LM을 fine-tuning 하기 위한 adapter 방법



Unlikelihood Training^{2 3}

- 생성을 지양하는 token 혹은 sequence에 대해 likelihood training의 역방향 학습

- Likelihood Training: $\mathcal{L}_{MLE} = -\sum \log p_{\theta}(y_t | x, y_{<t})$

- Unlikelihood Training: $\mathcal{L}_{UL} = -\sum \sum_{y_c \in C} \log p_{\theta}(1 - (y_c | x, y_{<t}))$

¹ Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv preprint arXiv:2106.09685*(2021).

² Welleck, Sean, et al. "Neural Text Generation With Unlikelihood Training." *International Conference on Learning Representations*. 2019.

³ Li, Margaret, et al. "Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

2.3 Unlikelihood Training Example

AI: 안녕하세요. 클로바 케어콜입니다. 어르신 오늘 외출하실 일은 없으신가요?

User: 날씨가 좋아서 산책이나 나갈까 생각중이에요

AI: 그러시군요. 혹시 자주 가시는 곳이 있으신가요? ♀

User: 집근처 공원이요

AI: 그렇군요. 그런데 혼자 나가시기는 좀 위험하지 않으신가요? ♀

User: 그렇긴한데 딱히 갈데도 없고 해서 그냥 다녀요

AI: 그러시구나. 제가 같이 동행해드릴까요? ✕

User: 그래주면 고맙지요

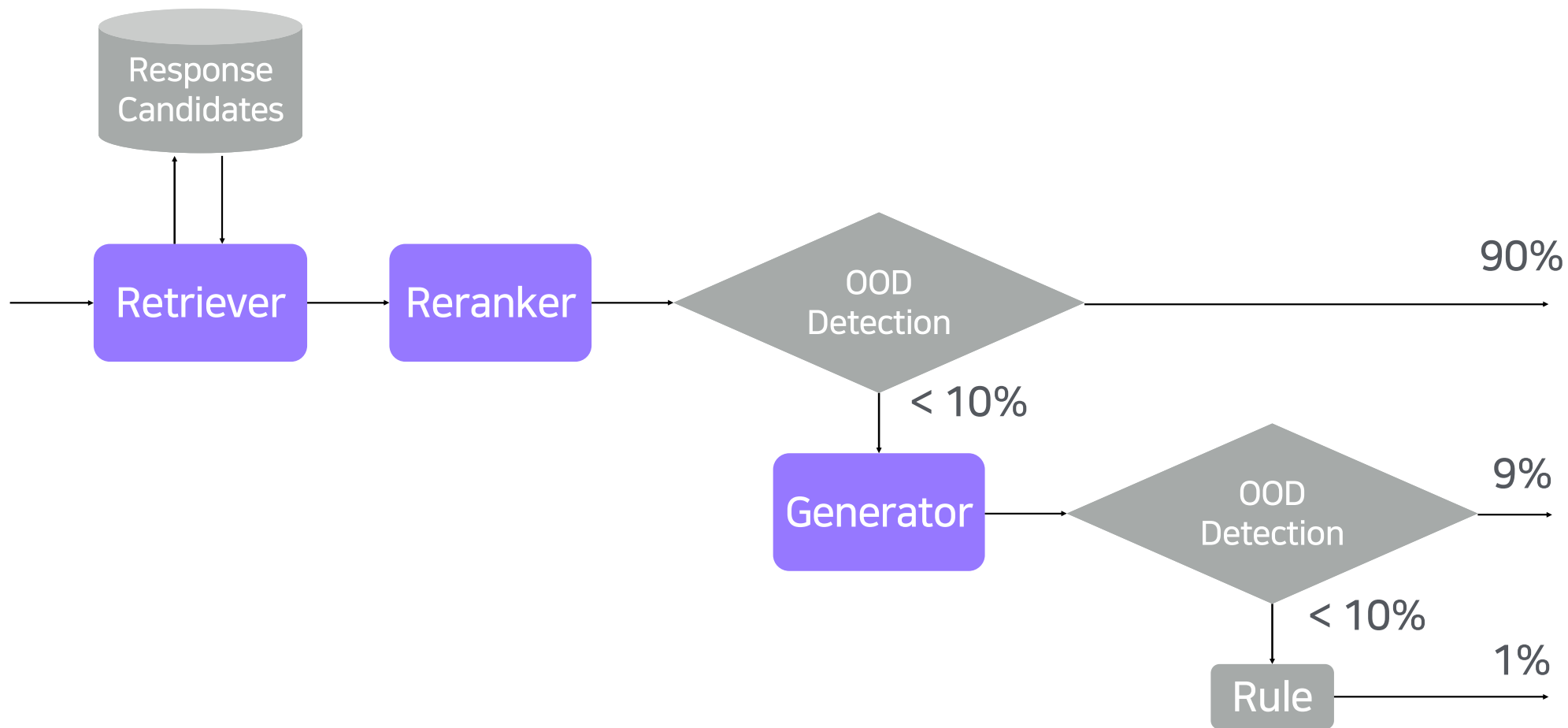
AI: 그러면 준비되시면 말씀해주세요. 바로 출발하겠습니다.

User: 알았어요

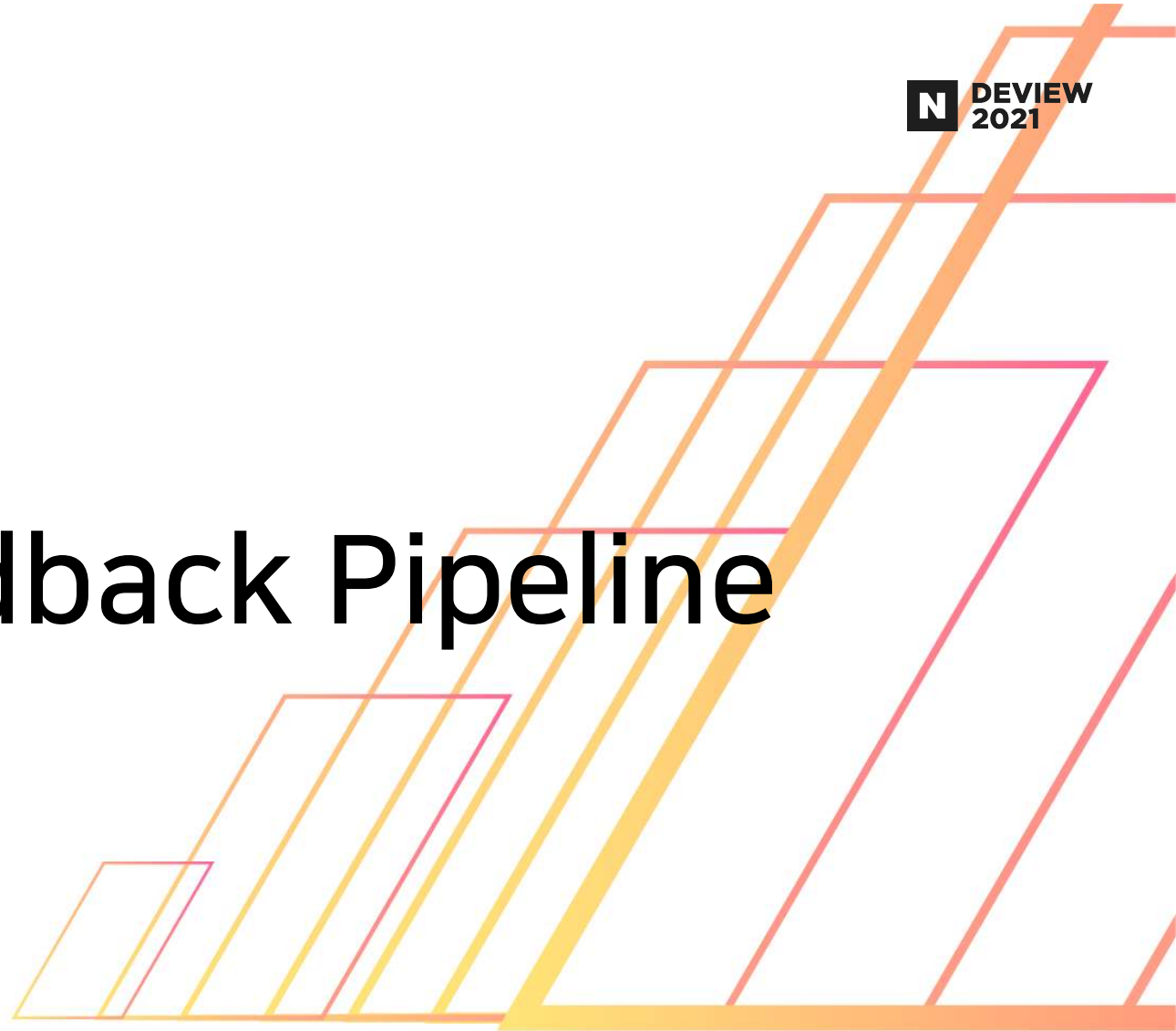
2.3 Generator Evaluation



2.3 Model Pipeline



2.4 Feedback Pipeline



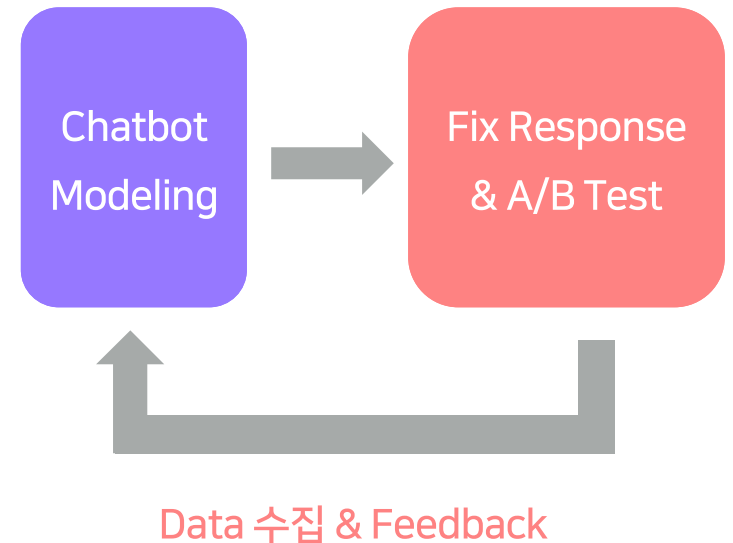
2.4 Human-in-the-Loop

사용자와의 대화를 통한 데이터 수집

- Retrieval 모델의 부족한 응답 후보 보완
- Generative 모델의 positive/negative example 추가

데이터 수집과 평가 동시 수행

- 수집된 데이터를 이용한 재학습
- 평가 지표 활용 및 모델링 반영



2.4 Fix Response

잘못된 응답 수정하기

- Chatbot Agent와 작업자의 대화
- Fix Response 버튼 → HyperCLOVA를 통한 대안 발화 생성
- Positive example (수정된 대화), negative example (수정전 대화) 수집

Dialogue Context

AI: 어르신, 안녕하세요~ 요새 별일 없으세요?
 User: 네. 별 일 없어요.
 AI: 다행이네요. 저는 얼마 전에 친구들이랑 여행을 다녀왔어요
 User: 어디로요?
 AI: 제주도요!
 User: 제주도 좋죠. 가서 뭐했어요?
 AI: 오랜만에 바다도 보고 맛있는 것도 많이 먹었어요.
 User: 뭐 먹었나요?
 AI: 흑돼지랑 갈치조림이요! {HyperClove}
 User: 오 좋았겠네요.
 AI: 근데 사람이 너무 많아서 좀 힘들었어요
 User: 아 요즘에 제주도에 사람이 많죠.
 AI: 그럼요 그래서 그런지 사람들이 엄청 많더라고요
 User: 그래도 저도 가고 싶네요.
 AI: 나중에 시간되시면 가족분들이랑 함께 다녀오세요~

Choose from predefined subjects

근황1

Send Message

insert message

2.4 A/B Test

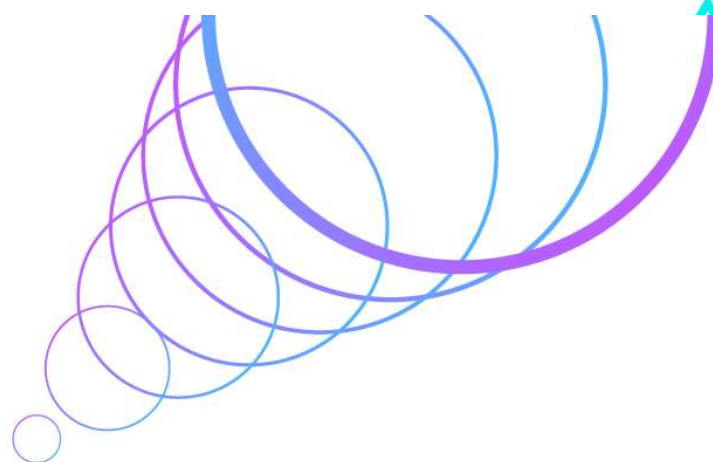
Fix Response 확률 \approx 잘못된 응답 발생 확률


- NAVER CLOps (CLOVA + MLOps)의 A/B 테스트 기능 활용
- 각 비교 모델이 50% 확률로 응답
- 정량적 지표 활용 및 모델 비교

- A** AI: 안녕하세요. 클로바 케어콜입니다. 요즘 운동은 꾸준히 하고 계신가요?
User: 나야 매일 하지 아침마다 공원가서 산책하구 저녁에는 헬스장 가서 운동한다우
- B** AI: 아 그러시군요. 건강관리를 위해 열심히 하시네요.
User: 내가 나이가 있다보니 몸이 예전같지 않아서 말이야 그래서 이렇게 툼툼히 운동을 해주고 있지
- B** AI: 그렇군요. 혹시 좋아하는 운동이나 잘하는 운동이 있으신가요?
User: 나는 당구를 좋아해 잘치지는 못하지만 친구들이랑 내기할때 종종 치곤했지
- A** AI: 오 당구도 좋아하시는군요. 언제 한번 같이 치러 가실래요? **X**
AI: 그러셨군요. 저도 예전에 친구들이랑 당구장에서 내기하면서 놀았던 기억이 나네요.

Fixed 

2.5 Demo: CareCall for Senior





3. HyperCLOVA로 문서 기반 QA Bot 만들기

3.1 Domain Specific Question Answering System

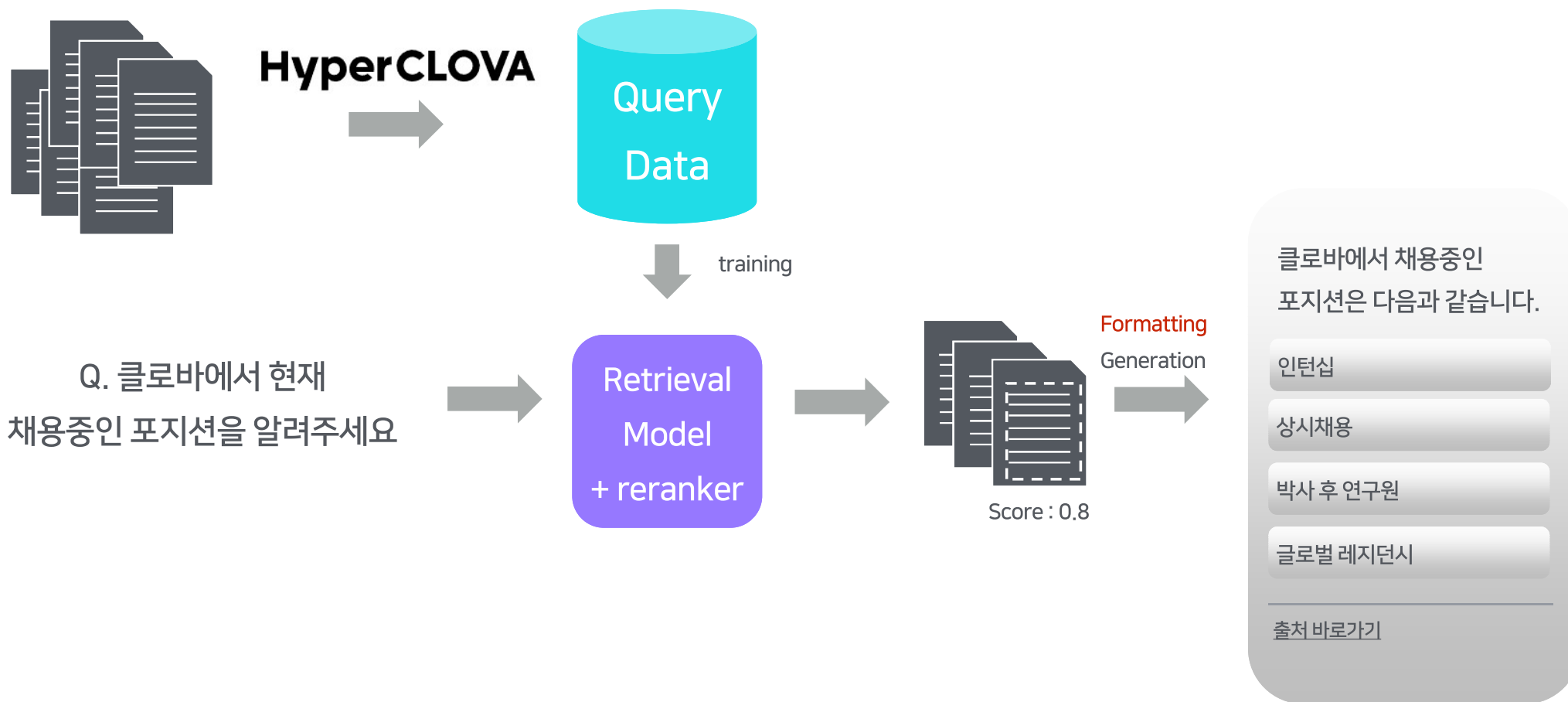
도메인 특화 QA 시스템을 위해

- 도메인에 특화된 모델을 위해서는 사람이 직접 생성한 데이터가 필수적이거나, 이를 HyperCLOVA로 대체하여 데이터 생성 비용을 대폭 감소

고려해야할 점들

- 도메인 내에만 존재하는 용어들
- 사용자의 전달방식이 달라질 수 있음
- Edge Project 로 clova.ai 홈페이지에 대한 정보 검색 기반 QA 시스템을 기획

3.1 Pipeline



3.2 Dataset Build



3.2 Data generation

Document-based Data Generation

- chit-chat 보다 out-case를 허용범위가 좁음
- Document 에서 답변 가능한 query, 내용과 일치하는 answer
- (Document , Query) → Retrieval Model 학습에 사용
- ((Query, Document), Answer) → Generator 학습에 사용

중국집 메뉴 정보
짜장면 -6000원
짬뽕 - 7000원
탕수육 - 10000원
...

생성 가능한 질문과 답변



질문 : 짜장면은 얼마인가요?
답변 : 6000원 입니다
질문 : 짬뽕은 얼마인가요?
답변 : 7000원 입니다
질문 : 탕수육도 파나요?
답변 : 네, 탕수육은 10000원에 팔고
있습니다

3.2 Query Generation

HyperCLOVA

Document Example 1

Query 1

Query 2

Document Example 2

Query 1

Query 2

Target Document

Query ?

생성 결과

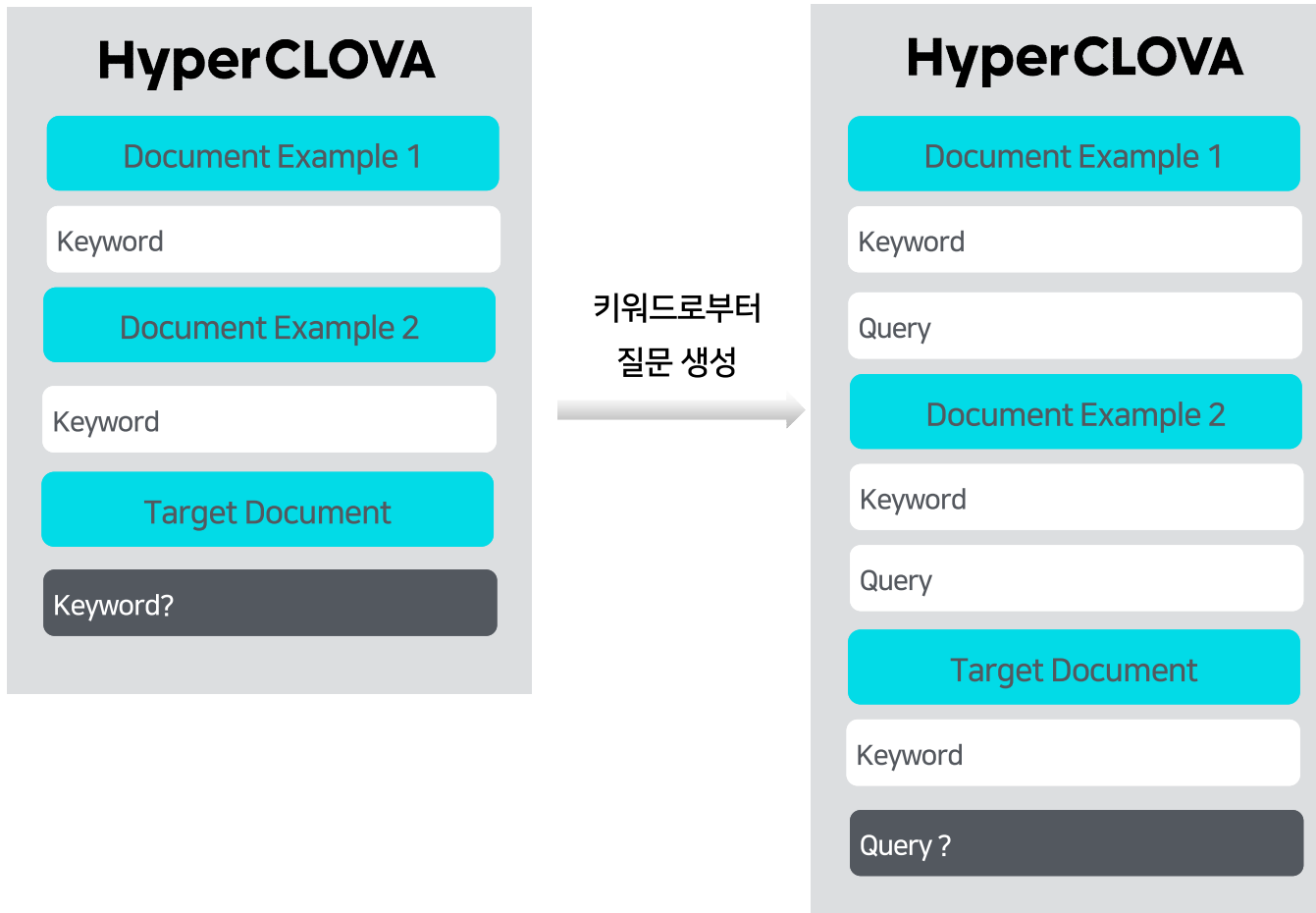
clova.ai 홈페이지 문서>리서치>Careers>상시채용>지원자격

- CLOVA의 성능향상을 위한 AI 선행 기술 연구가 가능한 인력 선행기술 연구 전문가 (Research Scientist)
- 최신 AI 연구 기술에 대한 이해, 우수한 연구 역량 및 실적 AI SW 엔지니어
- Tensorflow, PyTorch, MXNet, Caffe2 등 오픈소스 프레임워크 기반 개발역량
- 멀티 GPU 및 고성능 컴퓨팅관련 업무 경험

- Q. CLOVA Clock 프로젝트의 AI 선행기술 연구 전문가는 무슨 일을 하나요?
- Q. 어떤 일을 할 수 있어야 할까?
- Q. CLOVA Research 조직에서 하는 일은 뭐가 있어요?
- Q. AI 연구원 지원하려고 하는데 이 조건이 다 해당되나요?
- Q. CLOVA AI research는 어떤 일을 하는 팀인가요?
- Q. CLOVA AI research는 무엇을 하는 조직인가요?

다양한 쿼리가 생성되지 않고, 실제로 질문이 나올만한 부분이 아닌, 특정 단어에 얽매어 질문이 생성됨

3.2 Query Generation-키워드 기반 질문 생성



3.2 Query Generation-키워드 기반 질문 생성

문서 :

clova.ai 홈페이지 문서>리서치>Careers>상시채용>지원자격
- CLOVA의 성능향상을 위한 AI 선행 기술 연구가 가능한 인력
(자연어처리, 컴퓨터비전, 추천 등을 포함한 기계학습기반 AI기술 전 분야)

선행기술 연구 전문가 (Research Scientist)

- 최신 AI 연구 기술에 대한 이해, 우수한 연구 역량 및 실적
- 프로젝트를 리딩하고 연구방향을 제시할 수 있는 강력한 연구개발 리더십

AI SW 엔지니어

- Tensorflow, PyTorch, MXNet, Caffe2 등 오픈소스 프레임워크 기반 개발역량
- 멀티 GPU 및 고성능 컴퓨팅관련 업무 경험
- 머신러닝/딥러닝 모델 설계 및 구축 경험
- 최신 AI 논문에 대한 빠르고 정확한 구현 능력



키워드 : 지원 자격, 선행기술 연구 전문가, AI SW 엔지니어, 우대 사항



Keyword : 지원자격

Q. 학력이 중요한가요?

Q. 지원 자격이 어떻게 되나요?

Keyword : 우대 사항

Q. AI SW 엔지니어의 우대사항은 어떻게 되나요?

Q. CLOVA 선행 기술 연구 전문가 (Research Scientist) 의 우대사항은 무엇인가요?

Keyword : 지원 자격 상세 내용

Q. 지원 자격 중 가장 중요한 것은 무엇인가요?

Q. AI 선행 기술 전문가는 어떤 일을 하나요?

3.2 Dataset statistics

HyperCLOVA를 통해 확보한 데이터셋

- 750 documents 에 대해 약 30000개 정도의 (q, D) 확보
- Human-made evaluation set : 6000개 정도의 (q, D)

CareCall for Senior처럼 검수가 필요하진 않나요?

- noise가 있는 질문들이 retrieval model의 성능을 오히려 높여주는 것을 실험적으로 경험 (distant supervision)

3.3 QA System Modeling

3.3 Retrieval Model Pretraining

Retrieval Model

- Retrieval model : Poly Encoder^[1]
- CareCall: dialogue history → response, QA : query → Document
- 사내에서 활용하는 larva model* (한국어 BERT) 을 기반으로 시작

* 언어 모델 학습과 사용 쉽게 하세요! : 효율적이고 확장성 있는 사내 라이브러리 개발기 (DEVIEW 2021)

Reranker

- Reranker : Cross Encoder
- Retrieval 로부터 나온 20개의 문서를 다시 랭킹
- KLUE MRC 로 pretraining

[1] LSamuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In 8th International Conference on Learning Representations, ICLR.

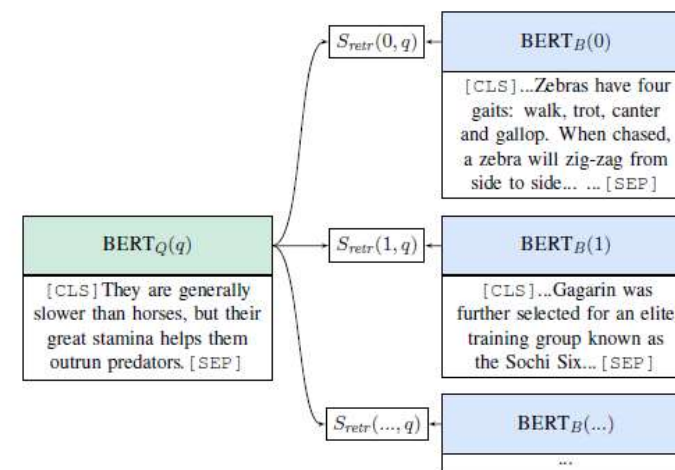
3.3 Retrieval Model Pretraining

Retrieval Model Pretraining Method

- Unsupervised Retrieval Pretraining Method
- Inverse Cloze Task (ICT)^[1], REALM^[2], ...

Inverse Cloze Task

- paragraph로부터 랜덤하게 문장 하나를 고르고, 나머지를 해당 문장에 대한 문서로 (q, D)
- Wikipedia 등 큰 corpus로부터 **unsupervised 하게 retrieval model을 pretraining 가능**
- 한국어 wiki 및 KLUE MRC 데이터로 pretraining
- 성능 향상 : q를 의문문 형태의 질문으로 변환



[그림] ICT 예시^[1]

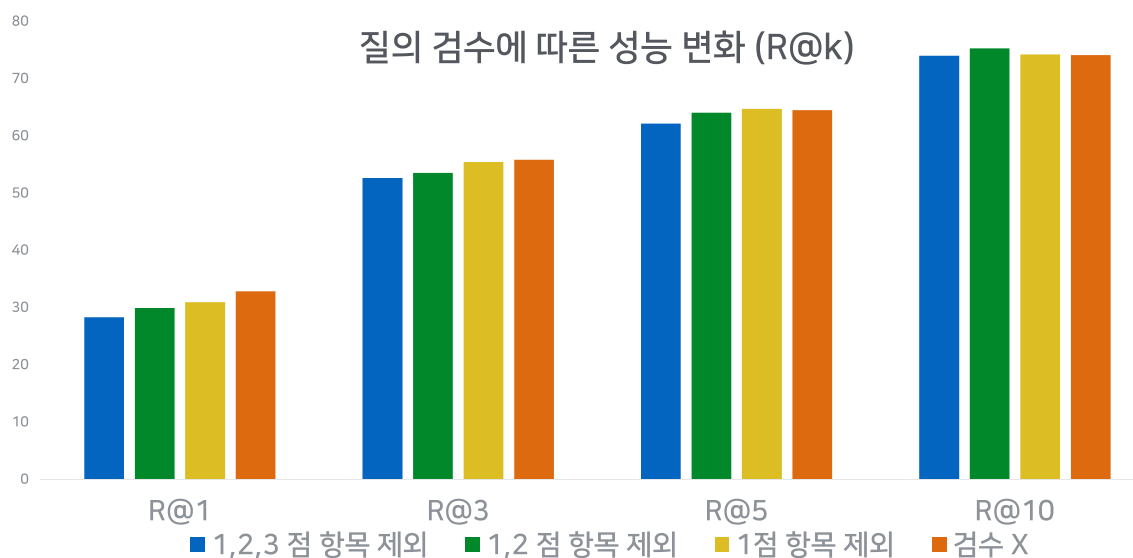
[1] Lee, K., Chang, M.-W., and Toutanova, K. Latent retrieval for weakly supervised open domain question answering. In Proceedings of the Conference of Association for Computational Linguistics, 2019

[2] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval augmented language model pre-training.

3.3 Retrieval Model Fine-tuning

데이터 검수에 따른 모델 성능 차이

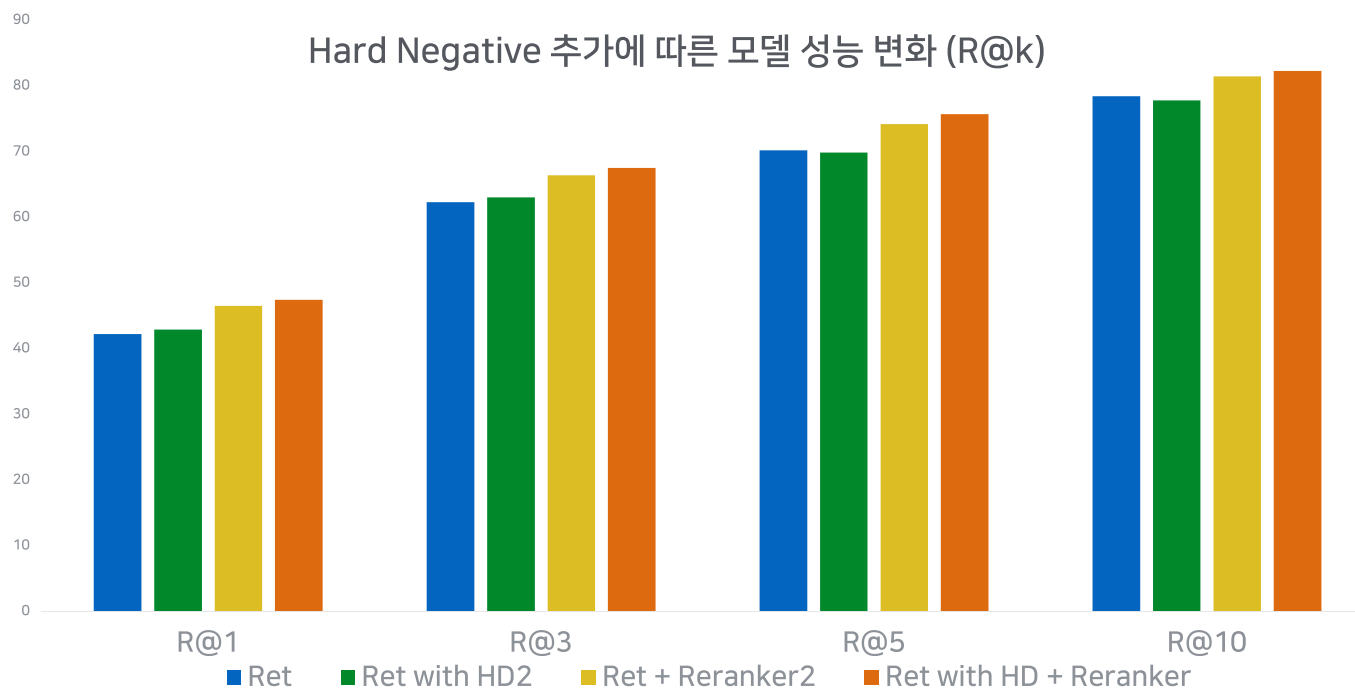
- 퀄리티에 따라 1~4점으로 평가된 검수 데이터 (1만개) 로 모델을 fine-tuning
- Metric : Recall@K (ground truth 문서가 rank k 안에 있는지를 확인)
- 검수되지 않은 데이터가 포함될 수록 성능이 증가함



3.3 Retrieval Model Fine-tuning

학습 성능 개선 - Hard Negative

- Hard Negative : positive document 와 매우 유사한 document
- BM25 score를 바탕으로, 한 batch내에서 Hard Negative Documents를 추가



3.3 Retrieval Model Fine-tuning

학습 성능 개선 - 동의어, 유의어 정보를 활용한 augmentation

- 기존에 model은 domain specific한 **유의어 및 동의어가 포함된 질문을 잘 인식하지 못함**
- Input regularize 이외에 학습 과정에서 사용해보고자 하는 아이디어
- 학습 과정에서 50%의 확률로 질문에 있는 단어를 동의어 및 유의어로 치환

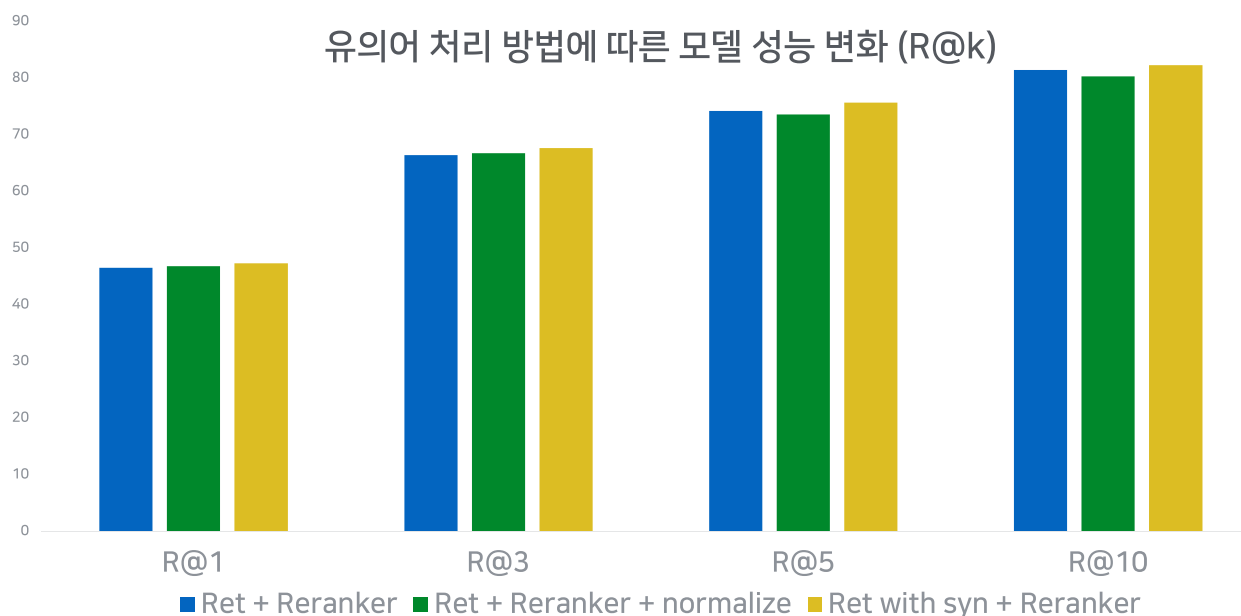
Clova : [클로바]

박사 후 연구원 : [포닥, Post-doc, 포스트닥터]

OCR : [광학 문자 인식, 오씨알]

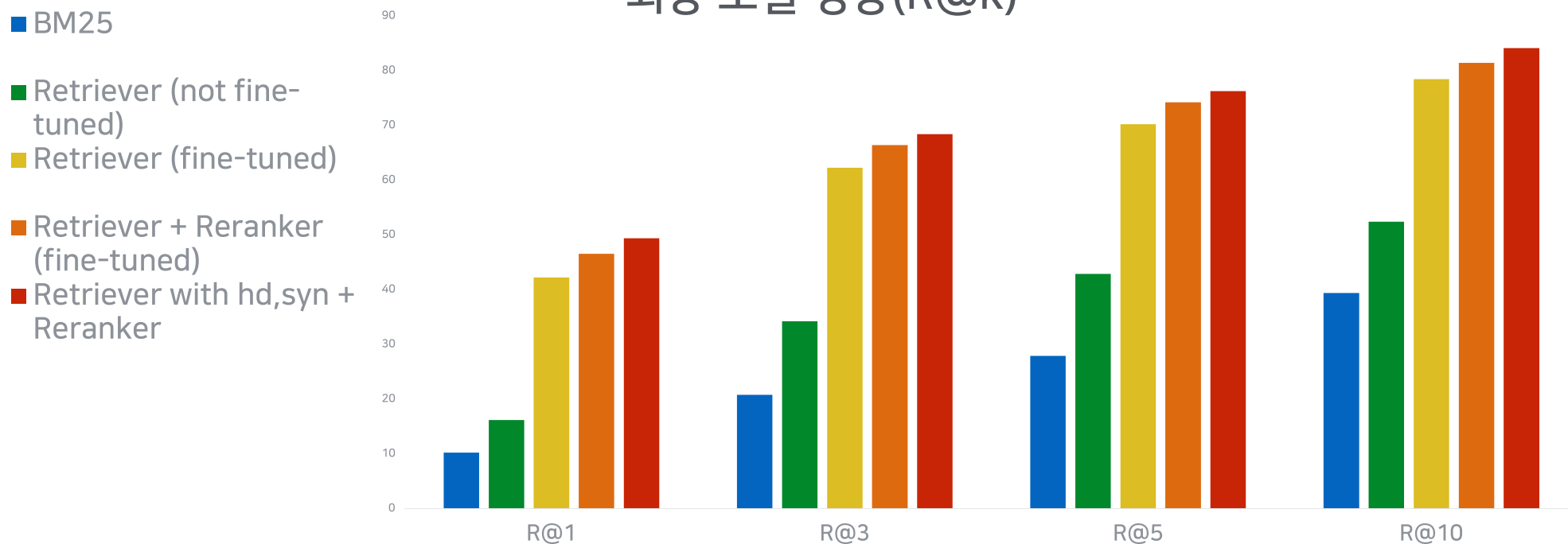
3.3 Retrieval Model Fine-tuning

학습 성능 개선 - 동의어, 유의어 정보를 활용한 augmentation



3.3 Retrieval Model Fine-tuning

최종 모델 성능(R@k)



3.3 Score Thresholding

모델로부터 나온 문서의 랭킹을 어떻게 활용할 것인가?

- 질문에 따라 활용해야하는 문서의 개수가 달라짐

Q. 클로바 램프 불빛 색상이 이상해요

- Top 3 문서가 필요

조명 색상에 따른 상태 표현

빨간색 불빛

노랑색 불빛

초록색 불빛

출처 바로가기

Q. 인턴 뽑을 때 우대조건이 있나요?

- Top 1 문서가 필요

인턴십 우대조건

- 영어 커뮤니케이션 역량

보유자

- 최신 AI 논문에 대한 빠르고

정확한 구현 능력(1~2주

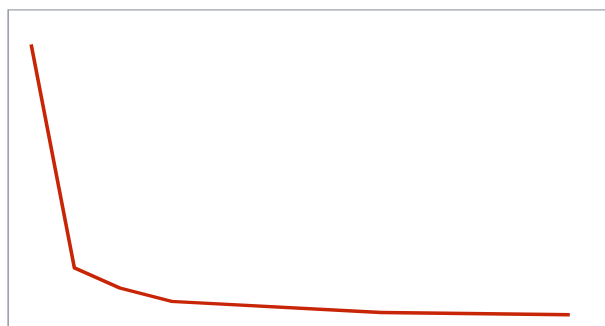
이내)

출처 바로가기

3.3 Score Thresholding

모델이 계산한 score를 확률 분포의 형태로 변환 (softmax)

- 확률 분포의 entropy를 바탕으로 Top1 및 Top3, OOD의 threshold를 조정



Top1 이 확실한 케이스
- entropy가 0에 가까움



Top3 정보가 모두 유용한 케이스
- Entropy가 0~1 사이

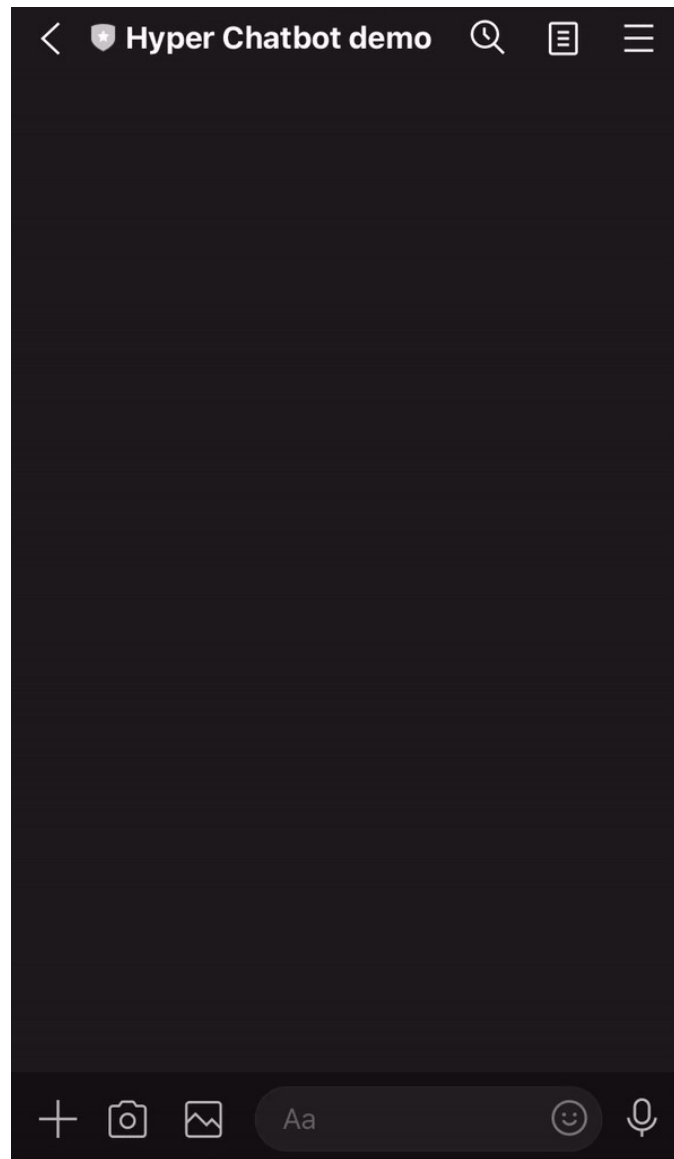


OOD (문서와 연관성이 하나도 없는 케이스)
- Entropy가 2 이상

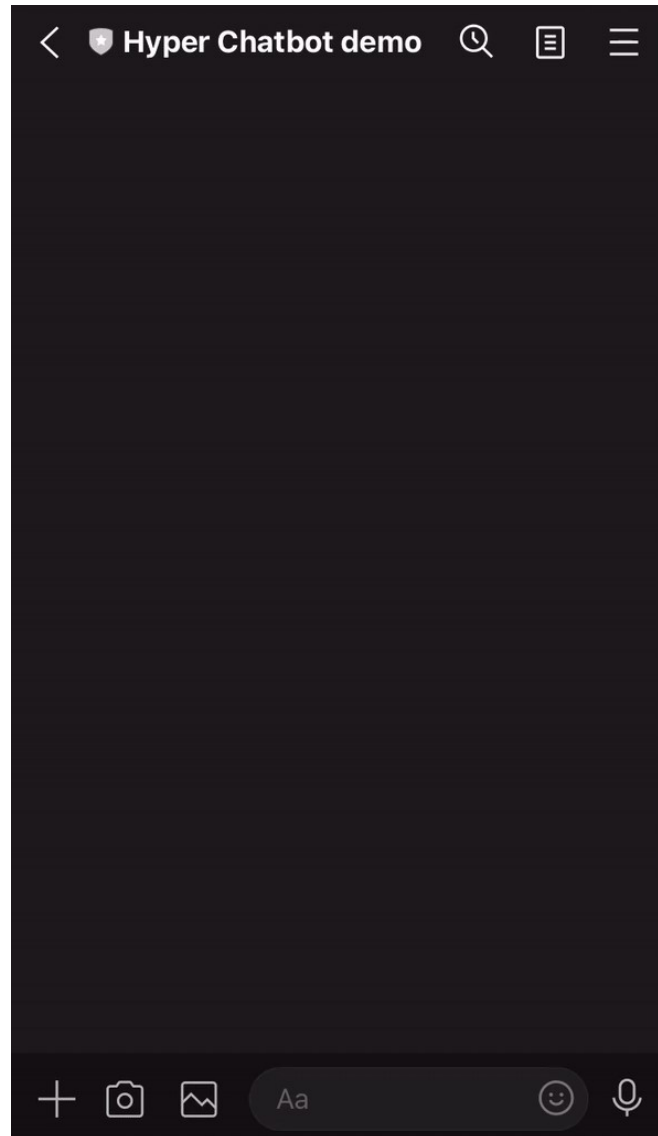
3.4 Demo



Top1 case

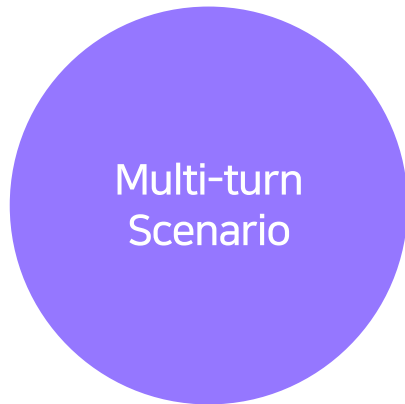


Top3 case



3.5 In Progress

3.5 현재 진행중인 사항들



Multi-turn 대화를 고려한 Retrieval

- Single turn : 유저의 목적이 한 턴에 모두 들어있다고 가정
- Multi-turn : 현재 턴에는 생략되었지만 (주어, 목적어) 위의 컨텍스트로부터 알 수 있는 정보들이 있을 수 있음
- 더 나아가 필요한 정보를 요청하고, 되묻는 능동적인 챗봇

단순 QA를 떠나 "대화"를 목표로 하는 방향성

3.5 현재 진행중인 사항들



Multi-turn
Scenario



Answer
Generation &
Filtering

Formatting Rule을 벗어나 모델이 답변을 생성

- 질문 생성보다 훨씬 복잡한 작업
- HyperClova로도 잘 답해주는 케이스도 있으나, Robust함을 우선시 해야함
- 사람의 검수가 무조건적으로 필요
- 문서 내에서 핵심 내용을 적은 문장으로 전달 가능

4. 서비스 및 앞으로의 계획

4.1 CareCall for Senior 서비스

해운대 구청을 통한 클로즈 베타 서비스

- 50 ~ 70대 어르신분들 대상으로 서비스
- 어려웠던 점 : 사투리 음성 인식, 인공지능전화 자체에 대한 낯설음
- 그럼에도 불구하고 전체적으로 긍정적인 평가



<https://teletobi-bbo.tistory.com/entry/해운대구청-홈페이지-바로가기>

- Conversation Planning 팀의 김혜리님, 박현정님께 감사의 말씀을 드립니다.
- 많은 도움을 주신 해운대 구청에도 감사의 말씀을 드립니다.

4.1 CareCall for Senior 향후 계획

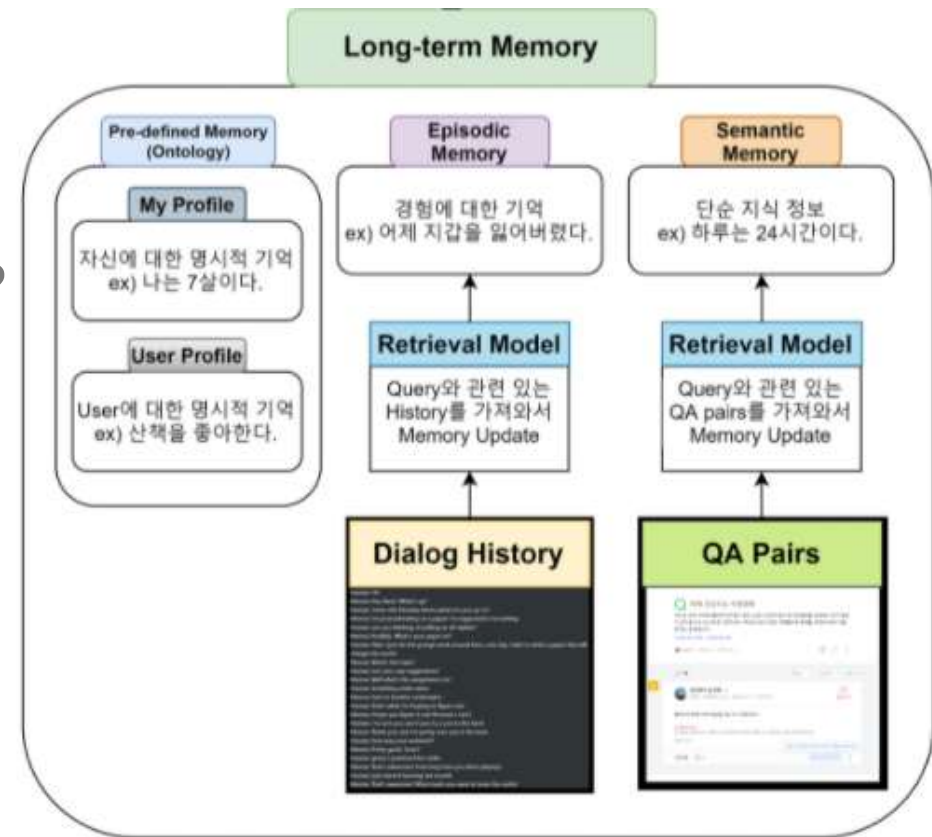
User State Tracking

- 유저의 상태 확인에 중요한 정보를 대화 내용으로부터 인식하는 모델
- 인식된 상태는 다시 대화에 활용됨
- 유저의 상태 변화를 감지하는 Tracking 및 추상적인 state에 대한 인식까지 확장 예정

4.1 CareCall for Senior 향후 계획

Long-term Memory

- 유저와 관계 형성을 위해서는 이전 대화에 대한 '장기 기억 능력' 이 필수
Ex) AI: 어르신 지난 주에 병원은 잘 다녀오셨어요?
- Retrieval와 Generator model이 과거 대화 기록을 활용하여 현재 문맥상 필요한 대화 수행
- 궁극적으로, 어떤 내용이 우선적으로 기억해야하는지 모델이 스스로 파악



4.2 문서기반 QA bot 서비스

연내 데모 서비스 예정

- 현재 내부 테스트 결과 만족스러운 사용자 체감 성능
- 글로벌 홈페이지의 공식 QA bot으로 공개하는 것이 목표

문서화 툴과 함께 QA봇 Framework의 확장

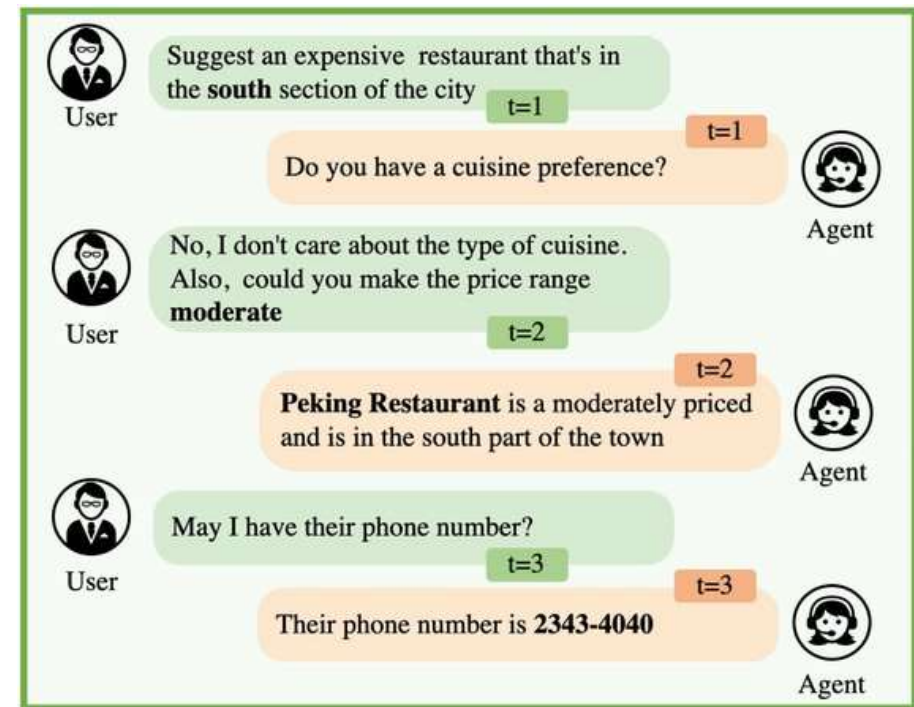
- 유저에게 보여줄 document를 정제하는 tool 제작
- 일반화된 QA봇 Framework Pipeline 개발

- Conversation Planning 팀의 김혜원님, 정유인님께 감사의 말씀을 드립니다.

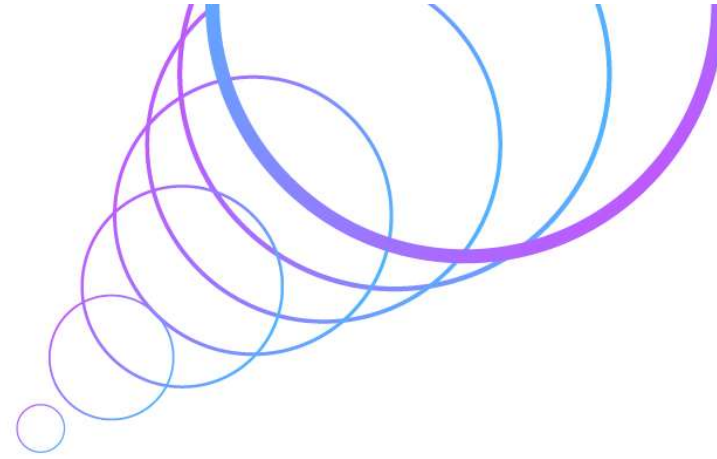
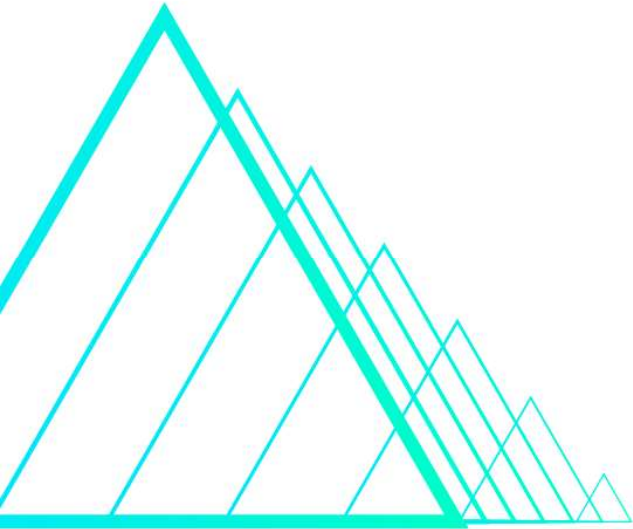
4.2 문서기반 QA bot 향후 계획

Task Oriented Dialog(TOD) 와의 결합

- 되묻기, 필요한 정보 질의 등 능동적인 회화 가능
Ex) 유저 : 와이파이 연결이 안되요.
 챗봇 : 어떤 제품을 사용 중이신가요?
- API call을 통한 실제 유저 request 처리 (제품 as 요청, 이벤트 참가 등록 등)



https://www.researchgate.net/figure/Example-of-a-training-task-oriented-dialog-At-turn-2-the-agent-first-queries-the-KB_fig1_351173284



Thank You

