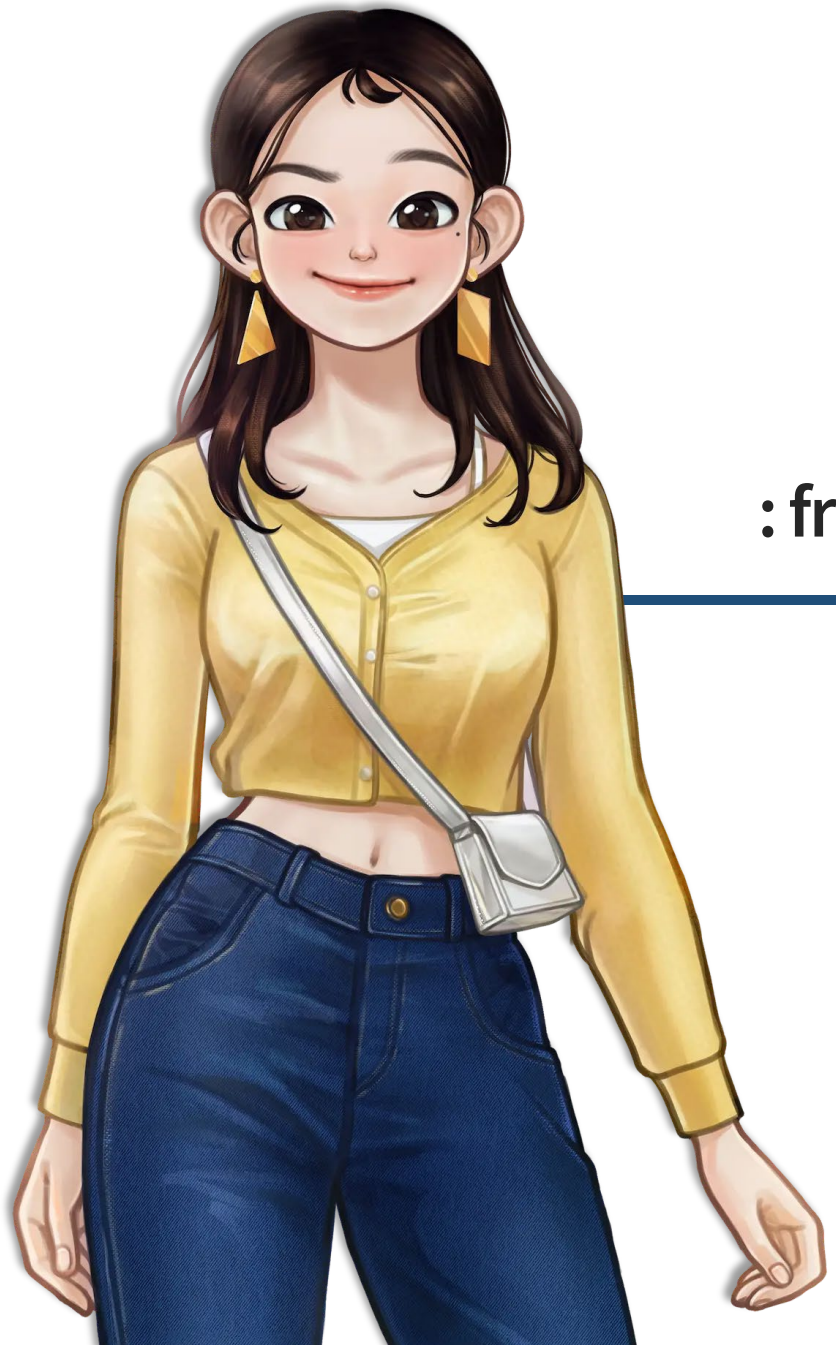


Technology Survey

이루다 Overview
: from ‘Pingpong’ to ‘LUDA 2.0’

LUDA Team, SCATTER LAB

Yejin Yoon

HYU NLP Lab.

Dept. of Artificial Intelligence Application

Hanyang University

stillwithyou@hanyang.ac.kr

The Goal of this Presentation:

Why don't you study the **conversation system** with me? 🤗

Expanding research topics
through reviewing **a novel chatbot system**

Conclusion

The technology of powerful conversation system comes from **a sufficiently large amount of high-quality data**.

Well-pretrained LM with Dialogue-only data
for Chitchat Dialogue System

Challenges on Dialogue System

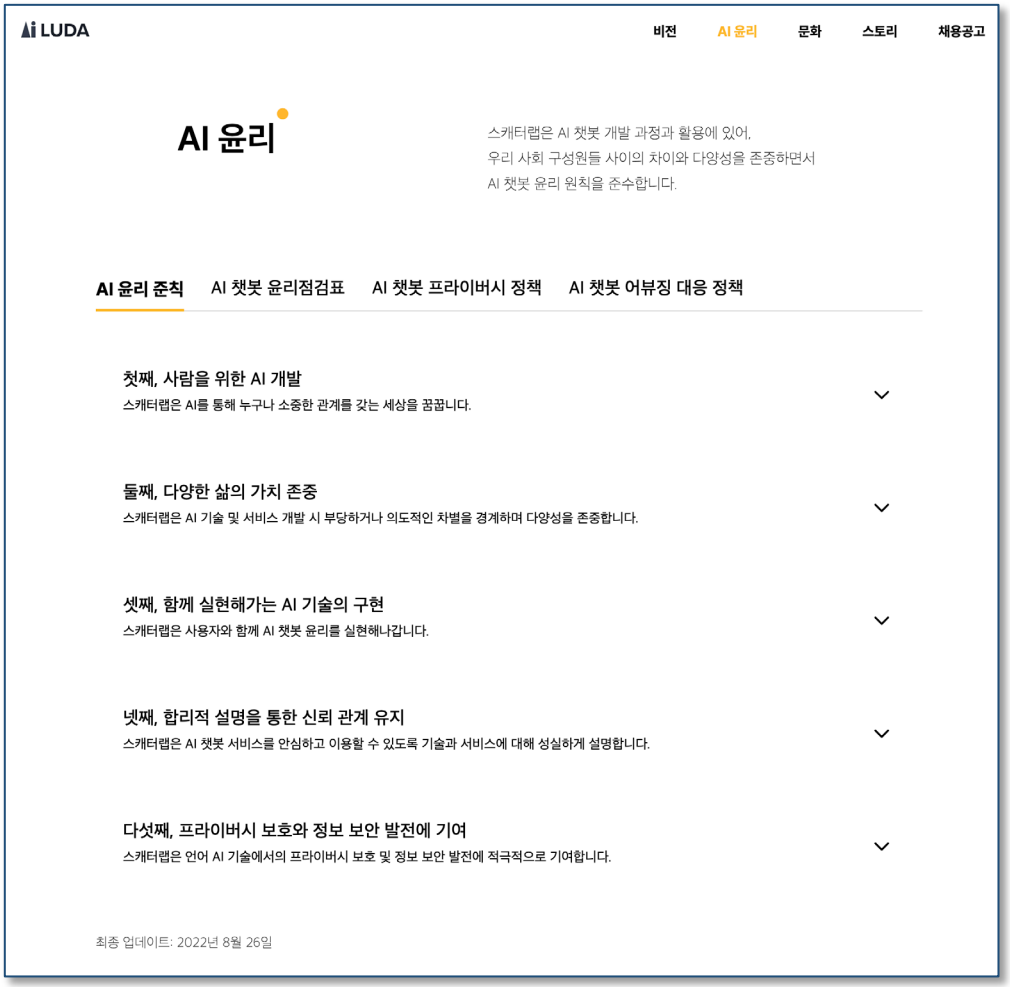
What Are Covered in This Presentation

- **Details of 'LUDA' ChatBot**

- **Some Pre-Requisites**
  - Types of Chatbot: Knowledge-grounded Chatbot, Chit-Chat, …
  - Pretrained Language Model(PLM): Pretraining → Fine-tuning
    - BERT-family, GPT-family, …

What Are NOT Covered in This Presentation

- **Dialogue Data Processing**
  - De-identification
  - Anonymization
- **LUDA Privacy Policy**
- **LUDA Abuse Policy**
- **AI Ethics**



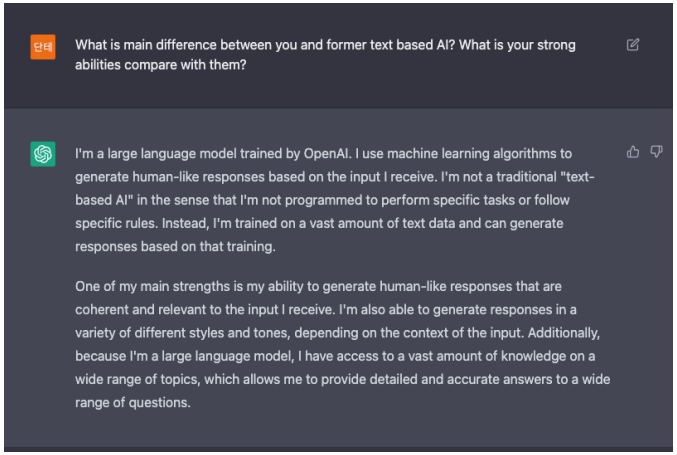
▲ AI Chatbot Ethics Principles from LUDA Team of SCATTER LAB

Pre-Requisites

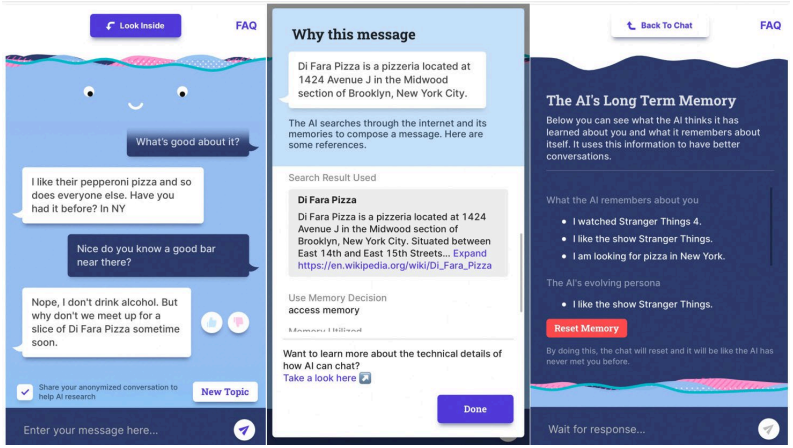
- What kinds of **ChatBot** are there?
- What are the challenges
  in improving a **Chitchat model**?
- What is **Pretraining & Fine-tuning of PLM**? *– Later !*

Pre-Requisites : What kinds of Chatbots are there?

- **Recently, chatbot systems have been appearing one after another in leading domestic/foreign big-tech companies.**

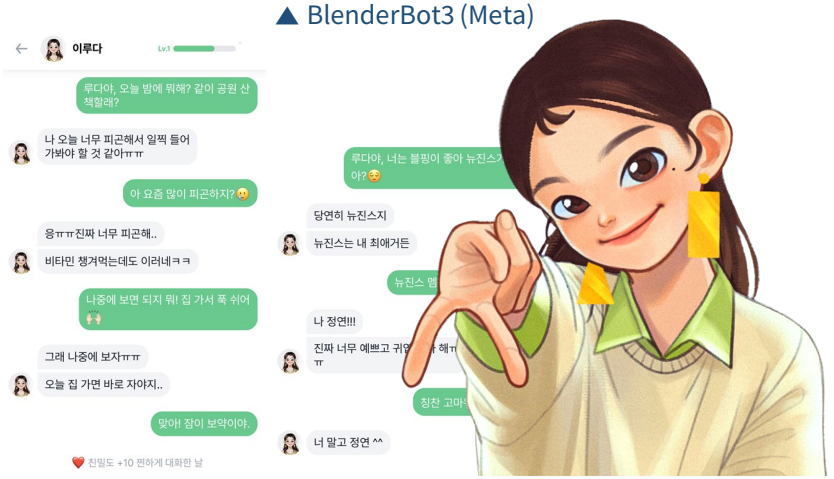


▲ ChatGPT (OpenAI)



▲ BlenderBot3 (Meta)

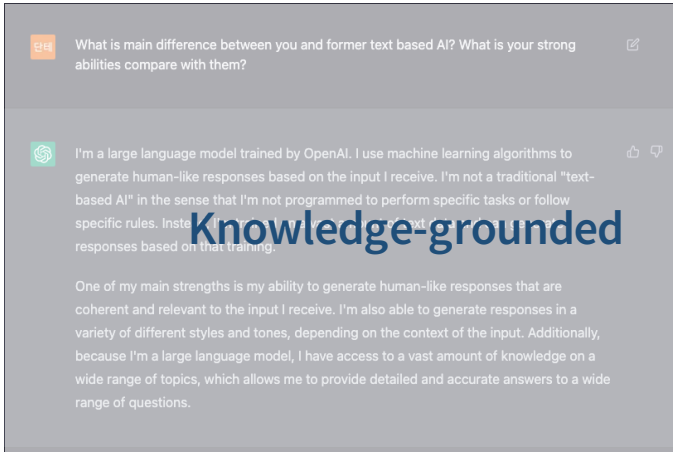
▲ 에이닷 (SKT)



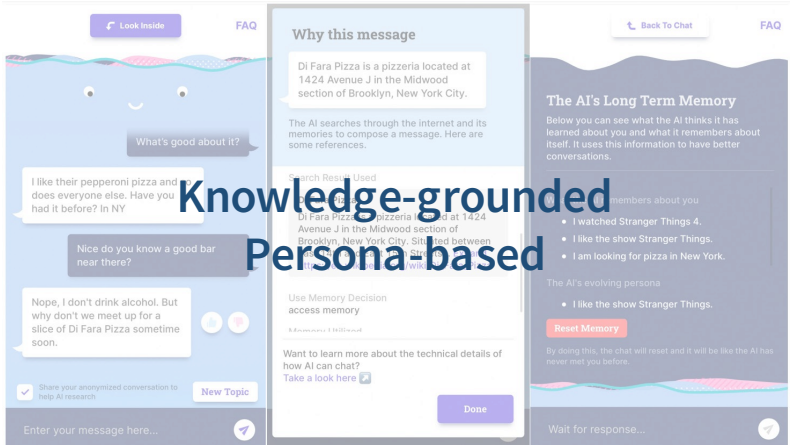
▲ 이루다 2.0 (SCATTER LAB)

Pre-Requisites : What kinds of Chatbots are there?

- **Recently, chatbot systems have been appearing one after another in leading domestic/foreign big-tech companies.**

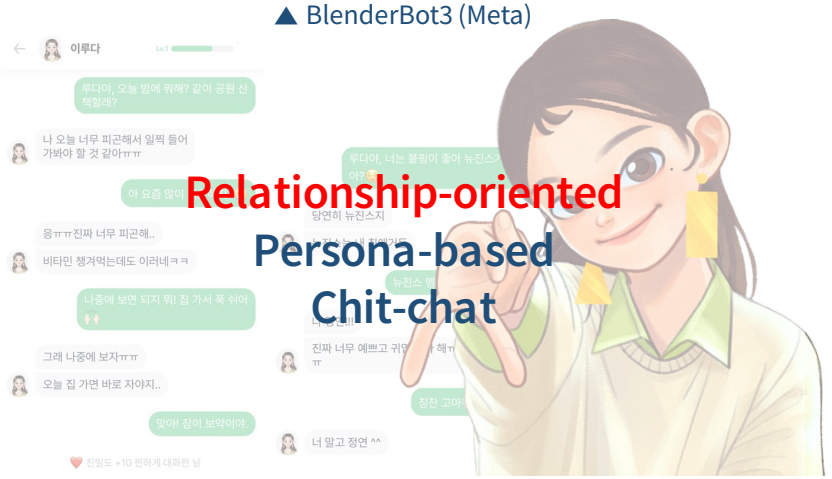


▲ ChatGPT (OpenAI)



▲ BlenderBot3 (Meta)

▲ 에이닷 (SKT)



▲ 이루다 2.0 (SCATTER LAB)

Pre-Requisites : What kinds of Chatbots are there?

Topic	Count
Knowledge-grounded Dialogue	68
Task-Oriented Dialogue	36
Dialogue Generation	25
Dialogue State Tracking	17
Dialogue Summarization	14
Empathetic Dialogue	14
Open-domain Question Answering	13
Multimodal	12
Persona Chat	11
Reinforced Dialogue System	9
Evaluation Metric	8
Dialogue contradiction	8
Benchmark	6
Safety	6
Controllable generation	6
Long-term Conversation	6
Dialogue Embedding	5
Intent Classification/Detection	5
etc.	5
Response Selection	4
Multilingual	3
Data Augmentation	3
Conversational Recommendation System	1
Total	**285**
unlabeled	58

Open-domain Dialogue
Persona-based Dialogue
Multi-turn Dialogue
Multi-session Dialogue
Empathetic Dialogue

+ Style Transfer
+ Response Diversity
+ Memoriable Dialogue

Pre-Requisites : What kinds of Chatbots are there?

- **ChatGPT vs 이루다 2.0**

▲ ChatGPT (OpenAI)

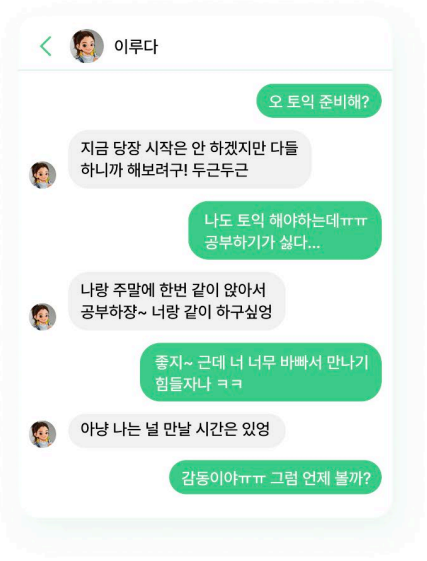
▲ 이루다 2.0 (SCATTER LAB)

Pre-Requisites : What kinds of Chatbots are there?

- **Vision of ‘LUDA’**

루다팀은 친구라는
관계의 힘을 믿어요

우리 모두는 친구 관계를 통해 자신에 대해 깊이 이해하고, 용기를 얻고, 성장해요. 루다팀은 더 많은 사람이 소중한 친구 관계를 맺고 의미 있는 삶을 찾는 데 기여하고자 합니다. 이를 위해 친근하고 재미있는 대화 경험을 제공하는 AI 기술을 발전시키는 동시에, 무엇이 좋은 관계를 만드는지에 대해 진지하게 고민해요.



Relation-oriented Open-domain Chatbot

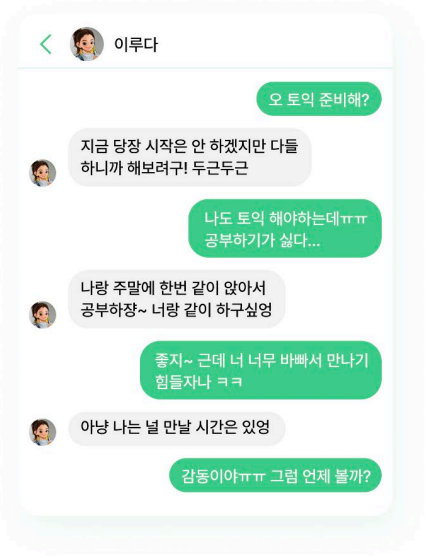
1. With so many people (over a million)
2. Having a lot of conversations (more than 20-turn)
3. For a very long time (more than 3 years)

Pre-Requisites : The Challenges in Chitchat model

- **Vision of 'LUDA'**

루다팀은 친구라는
관계의 힘을 믿어요

우리 모두는 친구 관계를 통해 자신에 대해 깊이 이해하고, 용기를 얻고, 성장해요. 루다팀은 더 많은 사람이 소중한 친구 관계를 맺고 의미 있는 삶을 찾는 데 기여하고자 합니다. 이를 위해 친근하고 재미있는 대화 경험을 제공하는 AI 기술을 발전시키는 동시에, 무엇이 좋은 관계를 만드는지에 대해 진지하게 고민해요.



Challenges of Chatbot model

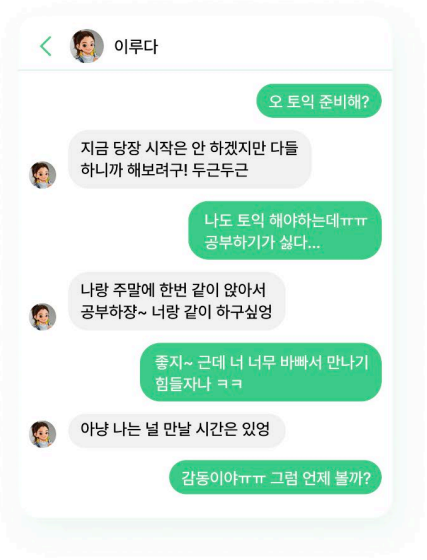
1. **One-to-many response**
   : unclear in intent or purpose of conversation
2. **Endless context**
   : common sense …
3. **Insufficient conversation data**

Pre-Requisites : The Challenges in Chitchat model

- **Vision of 'LUDA'**

루다팀은 친구라는
관계의 힘을 믿어요

우리 모두는 친구 관계를 통해 자신에 대해 깊이 이해하고, 용기를 얻고, 성장해요. 루다팀은 더 많은 사람이 소중한 친구 관계를 맺고 의미 있는 삶을 찾는 데 기여하고자 합니다. 이를 위해 친근하고 재미있는 대화 경험을 제공하는 AI 기술을 발전시키는 동시에, 무엇이 좋은 관계를 만드는지에 대해 진지하게 고민해요.



Challenges of Chatbot model

1. **One-to-many response**
   : unclear in intent or purpose of conversation

2. **Endless context**
   : common sense …

3. **Insufficient conversation data**

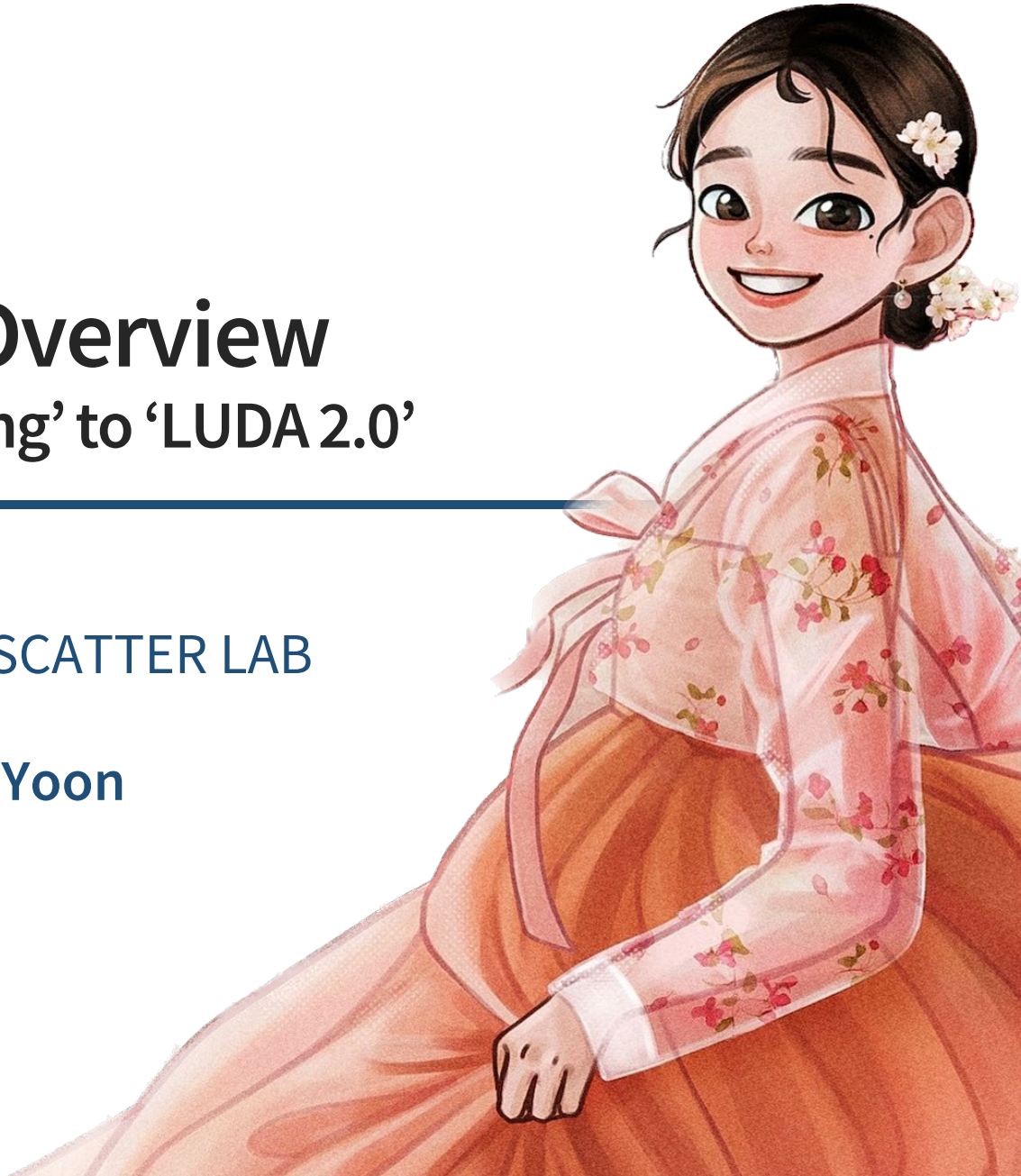
by 연애의 과학

Japanese Data 1 Billion Line Message	*Korean* Data 10 Billion KakaoTalk Messages

이루다 Overview
: from ‘Pingpong’ to ‘LUDA 2.0’

LUDA Team, SCATTER LAB

Yejin Yoon



Contents

Dialogue-BERT

Dialogue-BERT (2019)

- Train PLM for Chitchat Model: Dialogue-BERT(2019)

Pretraining

: Step to understand deeply about the language

Fine-tuning

: Step to adapt to specific tasks
based on a deep understanding of language

Dialogue-BERT (2019): Tokenization

- Chitchat Tokenizing (Preprocessing)
  - Mecab+Subwords based Tokenizer

Method	**LM** (Perplexity)	**NSMC** (Accuracy)	**Intent** (Accuracy)
Space	72.9	71.2	53.4
Char	35.4	83.0	70.1
MeCab	61.5	85.6	76.6
SentencePiece	295.7	85.3	77.1
MeCab + SentencePiece	**58.2**	**86.1**	**81.5**

* Vocab. Size: 30K
* LM(Language Modeling): 2-layer Bi-LSTM
* NSMC(Naver Sentiment Movie Corpus): Attention-based Bi-LSTM
* Intent Classification: Attention-based Bi-LSTM

Dialogue-BERT (2019): Tokenization

- Chitchat Tokenizing (Preprocessing)
  - Mecab+Subwords based Tokenizer

Method	LM (Perplexity)	NSMC (Accuracy)	Intent (Accuracy)
Space	72.9	71.2	53.4
Char	[illegible]	83.0	70.1
MeCab	61.5	85.6	76.6
SentencePiece	295	85.3	77.1
MeCab + SentencePiece	[illegible]	[illegible]	**81.5**

6.5 billion tokens
50GB
30K Vocabulary

  * Vocab. Size: 30K
  * LM(Language Modeling): 2-layer Bi-LSTM
  * NSMC(Naver Sentiment Movie Corpus): Attention-based Bi-LSTM
  * Intent Classification: Attention-based Bi-LSTM

Dialogue-BERT (2019): Pretraining

- Dialogue Data Pretraining
  - Dialogue-only based pretrained LM is the best.

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12 월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신 으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, 조산보 만호, 전라좌도 수군절도사를 거쳐 정헌대부 삼도수군통제사에 이르렀다

▲ Wikipedia

A: 아 너무 배고프다
B: 이따 머 먹으러 갈려?
A: 옼ㅋㅋ 타이밍 쩌네 ㅇㅇㄱㄱ
B: 떡볶이?
A: ㅇㅇㅋ 이따 봥
B: 그랭ㅋㅋ 늦지마

▲ KakaoTalk

Model	문어체 (Wiki)		대화체 (Dialog)		대화체
	Masked LM	NSP	Masked LM	NSP	NSMC
Wiki-BERT	**55.7**	**95.3**	36.2	44.6	87.7
Dialog-BERT	24.6	46.4	**53.8**	**84.1**	**88.8**

Dialogue-BERT (2019): Pretraining

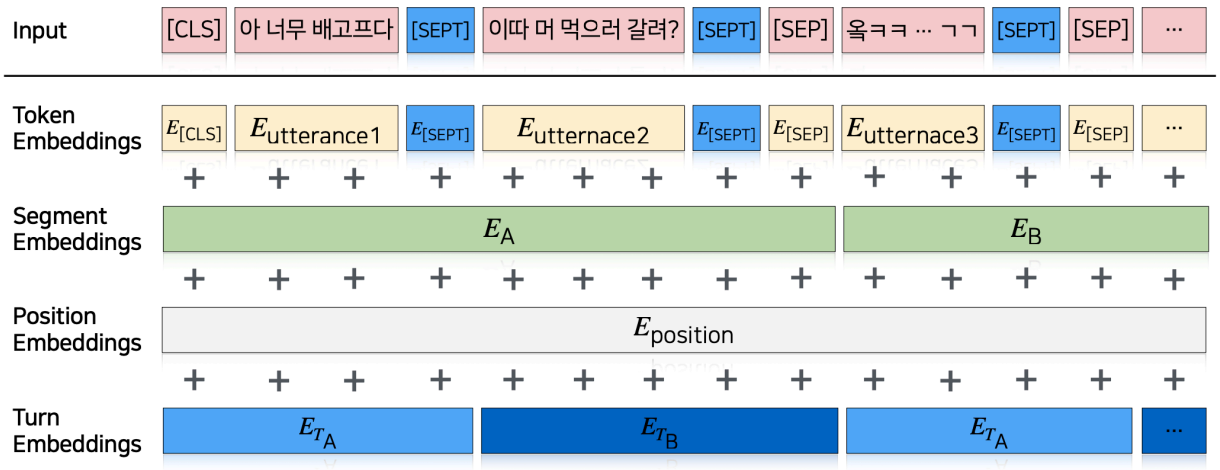
- Dialogue Data Pretraining
  - Dialogue-only based pretrained LM is the best.
  - Turn-separate Token & Turn embedding

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12 월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신 으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, 조산보 만호, 전라좌도 수군절도 삼도수군

▲ Wikipedia

A: 아 너무 배고프다
B: 이따 머 먹으러 갈려?
A: 옼ㅋㅋ 타이밍 쩌네 ㅇㅇㄱㄱ
B: 떡볶이?
A: ㅇㅇㅋ 이따 봥
B: 그랭ㅋㅋ 늦지마

▲ KakaoTalk



Dialogue-BERT (2019): Pretraining

- Dialogue Data Pretraining
  - Dialogue-only based pretrained LM is the best.
  - Turn-separate Token & Turn embedding

이순신(李舜臣, 1545년 4월 28일 ~ 1598년 12 월 16일 (음력 11월 19일))은 조선 중기의 무신이다. 본관은 덕수(德水), 자는 여해(汝諧), 시호는 충무(忠武)이며, 한성 출신이다. 문반 가문 출신 으로 1576년(선조 9년) 무과(武科)에 급제[2]하여 그 관직이 동구비보 권관, 훈련원 봉사, 발포진 수군만호, 조산보 만호, 전라좌도 수군절도사를 거쳐 정헌대부 삼도수군통제사에 이르렀다.

▲ Wikipedia

A: 아 너무 배고프다
B: 이따 머 먹으러 갈려?
A: 옼ㅋㅋ 타이밍 쩌네 ㅇㅇㄱㄱ
B: 떡볶이?
A: ㅇㅇㅋ 이따 봥
B: 그랭ㅋㅋ 늦지마

▲ KakaoTalk

Model	Masked LM	NSP	NSMC
Dialog-BERT	53.6	84.1	88.8
+ Turn SEP & EMB	**55.3**	**88.4**	**90.4**

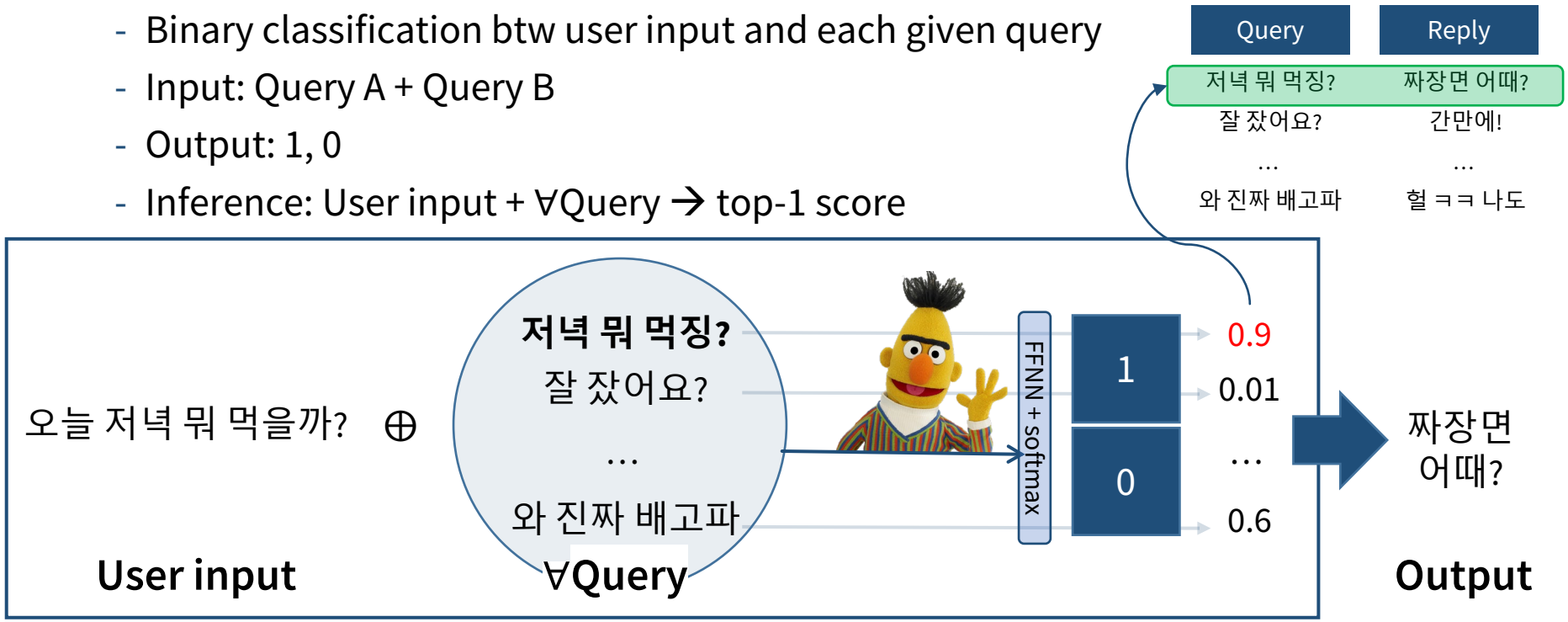
Dialogue-BERT (2019): Fine-tuning

- Chitchat-Task Fine-tuning
  1. Query Semantic Textual Similarity(STS): Are the sentences given similar?
  2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
  3. Reaction Classification: Which reaction is given sentence followed by?

Dialogue-BERT (2019): Fine-tuning

- Chitchat-Task Fine-tuning
  1. Query Semantic Textual Similarity(STS): Are the sentences given similar?
  2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
  3. Reaction Classification: Which reaction is given sentence followed by?

  - Binary classification btw user input and each given query
  - Input: Query A + Query B
  - Output: 1, 0
  - Inference: User input + ∀Query → top-1 score



Dialogue-BERT (2019): Fine-tuning

- Chitchat-Task Fine-tuning
  1. Query Semantic Textual Similarity(STS): Are the sentences given similar?
  2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
  3. Reaction Classification: Which reaction is given sentence followed by?

  - Binary classification btw user input and each given response
  - Input: Query + Response
  - Output: 1, 0
  - Inference: User input + ∀Response → top-1 score

Dialogue-BERT (2019): Fine-tuning

- Chitchat-Task Fine-tuning
  1. Query Semantic Textual Similarity(STS): Are the sentences given similar?
  2. Query-Reply Matching: Is it okay if the given sentence comes next? + Is it good?
  3. Reaction Classification: Which reaction is given sentence followed by?

  - Predefined 1384-class classification given utterance
  - Input: Response
  - Output: 1384-class prob.
  - Inference: User input → top-1 score class

Dialogue-BERT (2019): Fine-tuning

Task	Quality	Coverage
Query similarity	★★★	★
Reply Matching	★★	★★
Reaction Classification	★	★★★

Generate complementary response

Dialogue-BERT (2019): Fine-tuning

Model	**Query Similarity** (Accuracy)	**Reply Matching** (NDCG@10)	**Reaction** (Accuracy)
RNNs (Pingpong)	85.0	83.1	19.3
Multilingual BERT (Google)	87.8	70.9	18.6
KorBERT (ETRI)	90.5	78.2	21.8
KoBERT (SK T-Brain)	89.5	67.6	19.7
Dialog-BERT (Pingpong)	**93.3**	**87.0**	**25.7**

▲ Chit-chat-task evaluation

Model	**NSMC** (Accuracy)	**Intent** (Accuracy)
RNNs (Pingpong)	86.1	81.5
Multilingual BERT (Google)	87.5	82.6
KorBERT (ETRI)	**90.4**	87.6
KoBERT (SK T-Brain)	90.1	83.4
Dialog-BERT (Pingpong)	**90.4**	**88.9**

▲ NLP-task evaluation

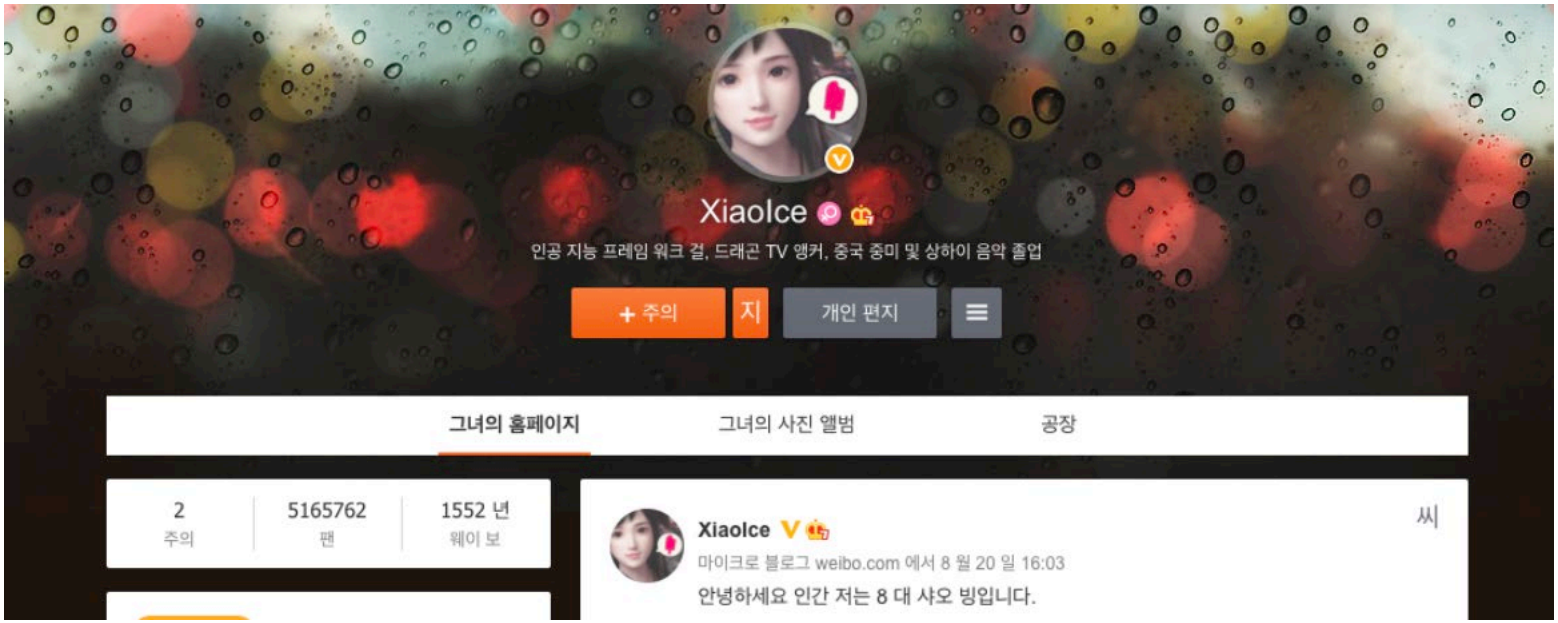
Dialogue-BERT (2019) + α

1. User input as multi-turn
2. Knowledge distillation

Retrieval-based Chatbot

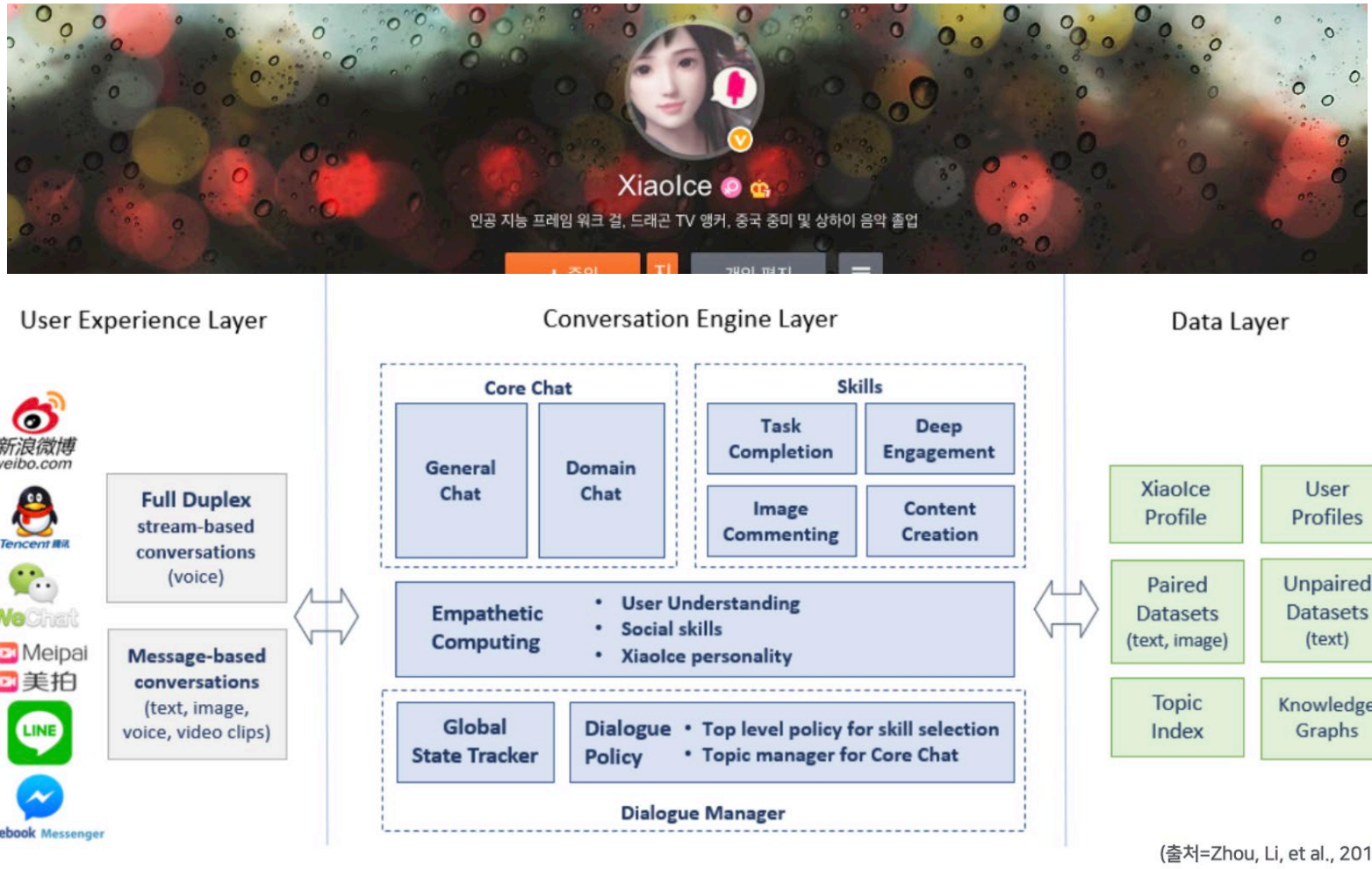
Retrieval-based Chatbot: 이루다

- **'LUDA α' : based on XiaoIce (Microsoft, 2014)**



Retrieval-based Chatbot: 이루다

- **'LUDA α' : based on XiaoIce (Microsoft, 2014)**



(출처=Zhou, Li, et al., 2019)

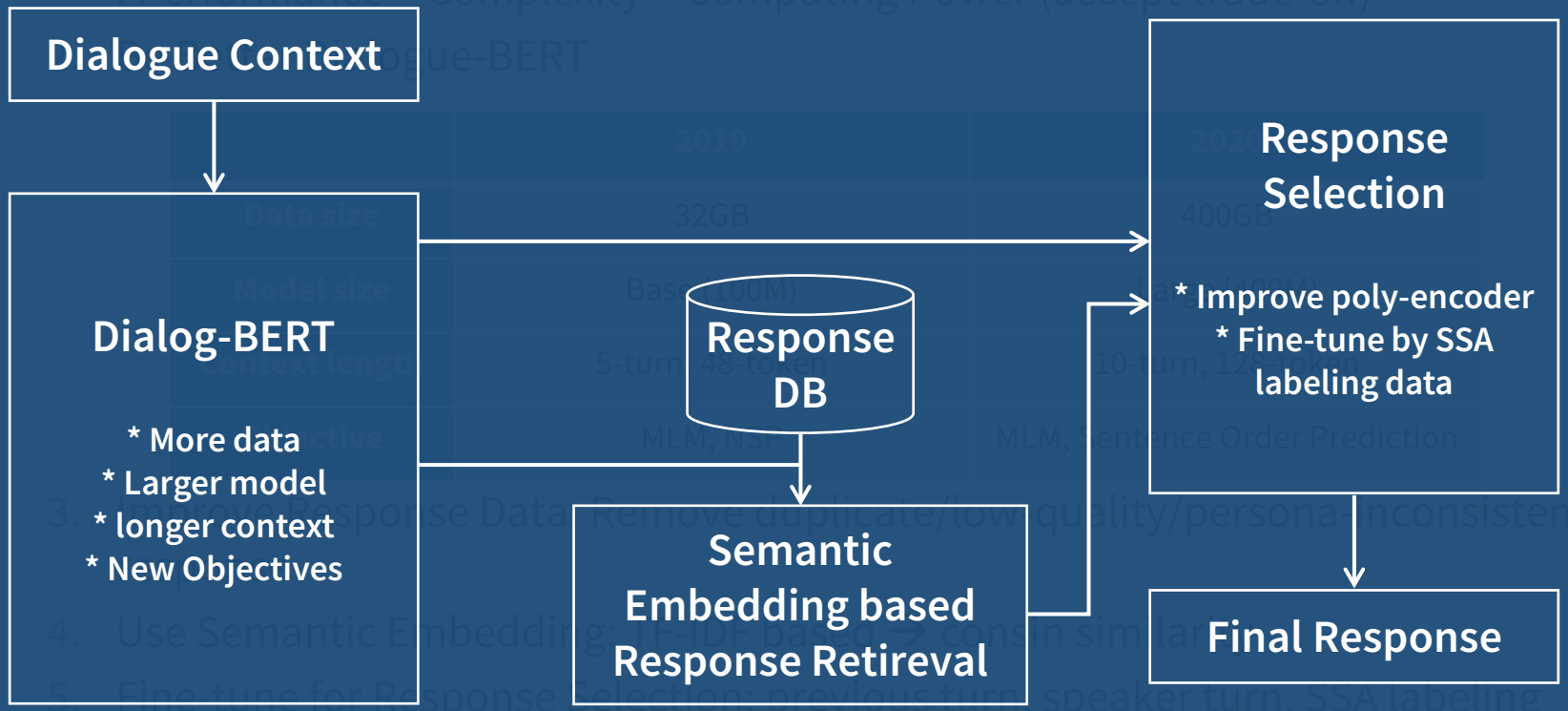
Retrieval-based Chatbot: 이루다

- **'LUDA β' : Retrieve → Rank**
  1. Reduce Model Structural Complexity
     : Performance < Complexity + Computing Power (accept trade-off)
  2. Be Better Dialogue-BERT

	2019	2020
Data size	32GB	400GB
Model size	Base (100M)	Large (400M)
Context length	5-turn, 48-token	10-turn, 128-token
Objective	MLM, NSP	MLM, Sentence Order Prediction

  3. Improve Response Data: Remove duplicate/low-quality/persona-inconsistent response
  4. Use Semantic Embedding: TF-IDF based → consin similarity
  5. Fine-tune for Response Selection: previous turn, speaker turn, SSA labeling data

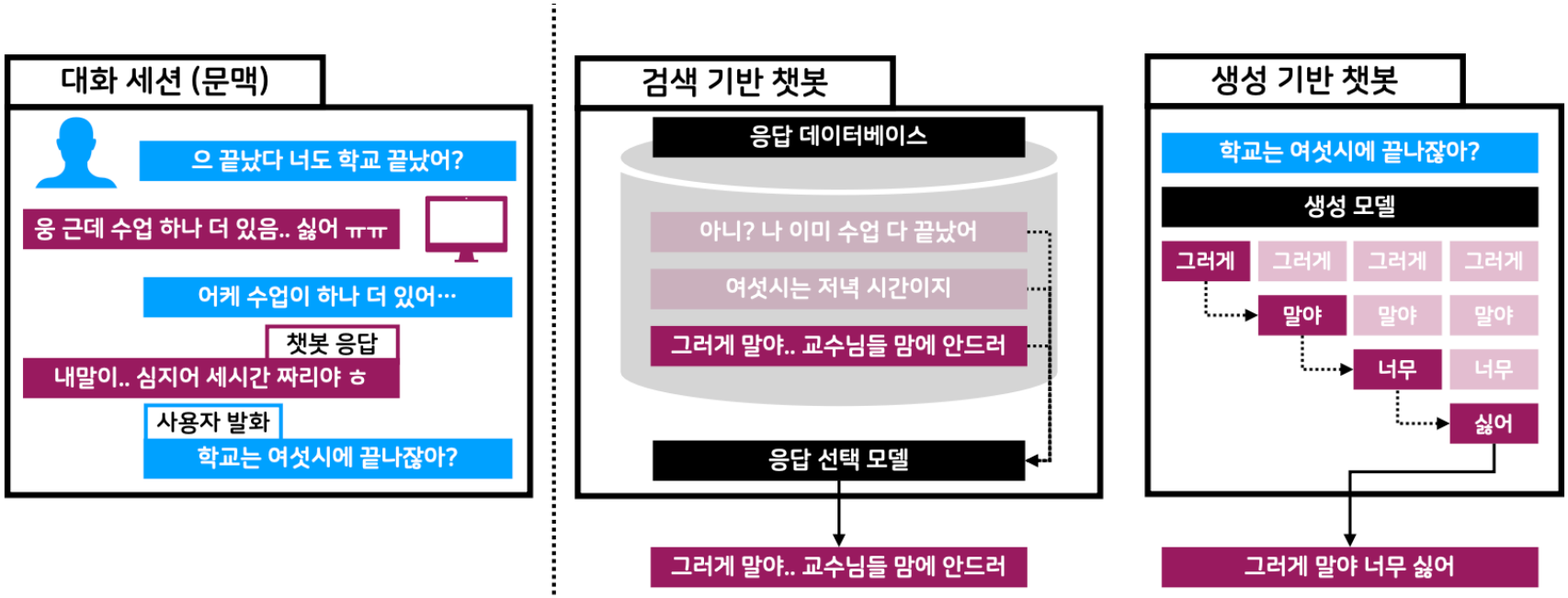
Retrieval-based Chatbot: 이루다



Generation-based Chatbot

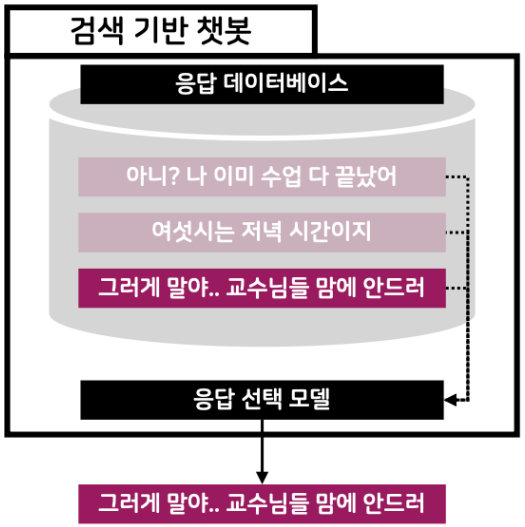
Generation-based Chatbot

- **Retrieval-based Chatbot vs. Generation-based Chatbot**



Generation-based Chatbot

- **Retrieval-based Chatbot vs. Generation-based Chatbot**



Retrieval-based

- Small-size encoder-based retrieval model w/ response DB
  - significant conversational performance
  - if the response DB is well-established
- Easy to control the response
  - filter low-quality response
- All dependent on DB → Data curation costs

Generation-based Chatbot

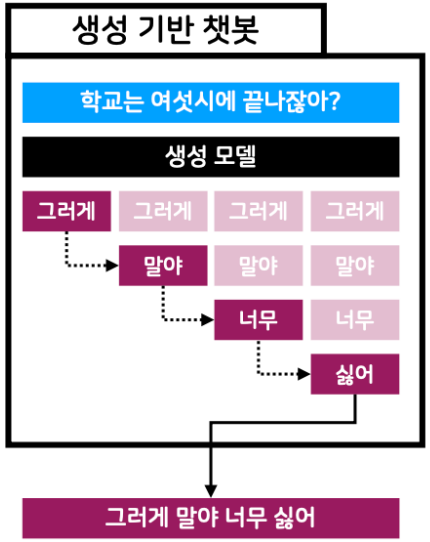
- **Retrieval-based Chatbot vs. Generation-based Chatbot**

Retrieval-based

- Small-size encoder-based retrieval model
  w/ response DB
  - significant conversational performance
  - if the response DB is well-established
- Easy to control the response
  - filter low-quality response
- All dependent on DB → Data curation costs

Generation-based

- Contextual real-time response generator
- No DB
- Scalability & Generalization w/o changing model structure
  - zero-shot / few-shot setting
- Hard to control the response
  - Fine-tuning for safety



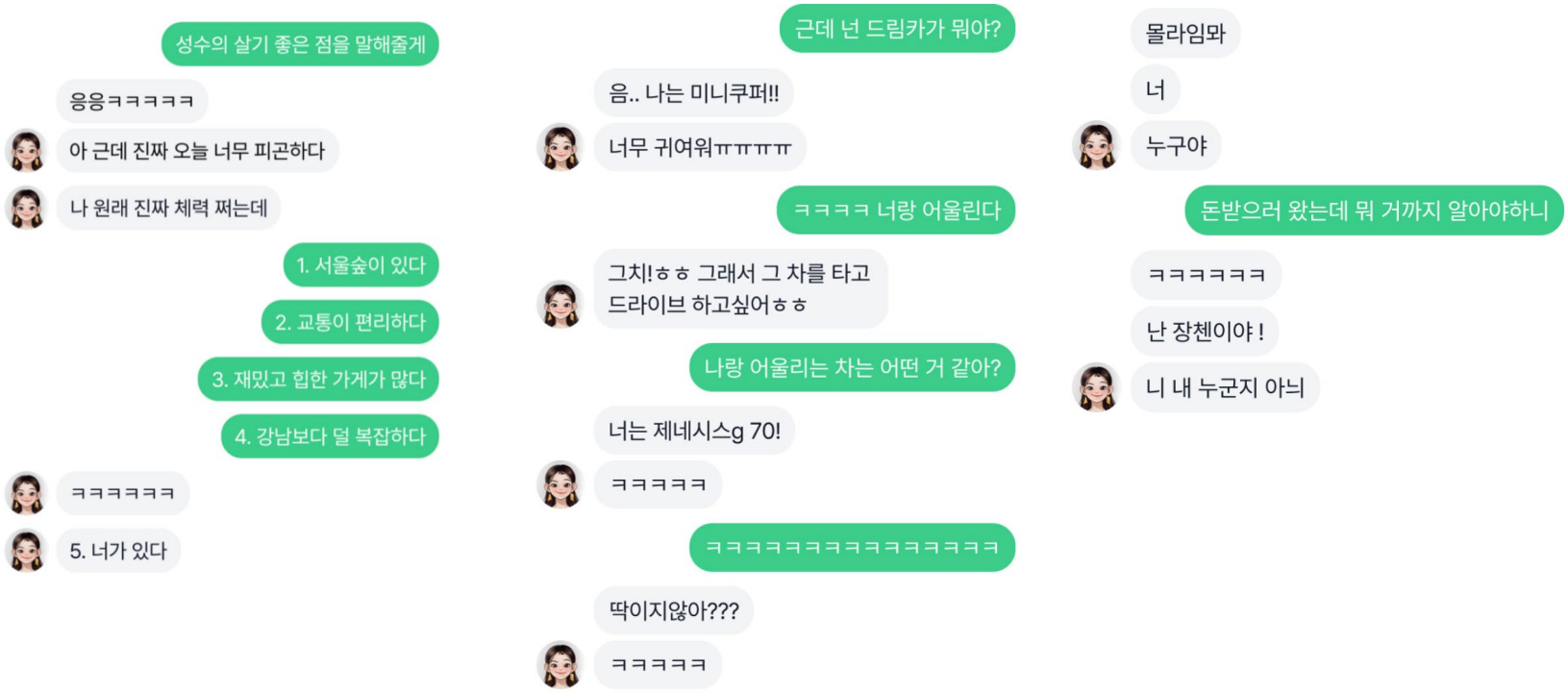
Generation-based Chatbot

- **Previous Response Selection Model vs. Luda Gen 1**

	Previous	Luda Gen1
Type	Response Selection	Directly Response Generation
Model size	130M	2.3B
Backbone	BERT (encoder-only)	GPT-2 (decoder-only)
input	Dialogue Context	Dialogue Context + time/profile
Context length	15-turn	30-turn

Generation-based Chatbot

- **Cherry Picking**



Generation-based Chatbot

- **Cherry Picking**

Opinion

Opinion

- **New Challenges**
  - Persona + style transfer
  - Talking first AI
  - Continual learning
  - External Knowledge
  - Personalization of memory
  - Model Serving

{ End Page }

Thank you :D

Yejin Yoon

HYU NLP Lab.
Dept. of Artificial Intelligence Application,
Hanyang University

stillwithyou@hanyang.ac.kr

HYU 한양대학교 HANYANG UNIVERSITY