

# Recipes for Building an Open-Domain Chatbot

백형렬

---

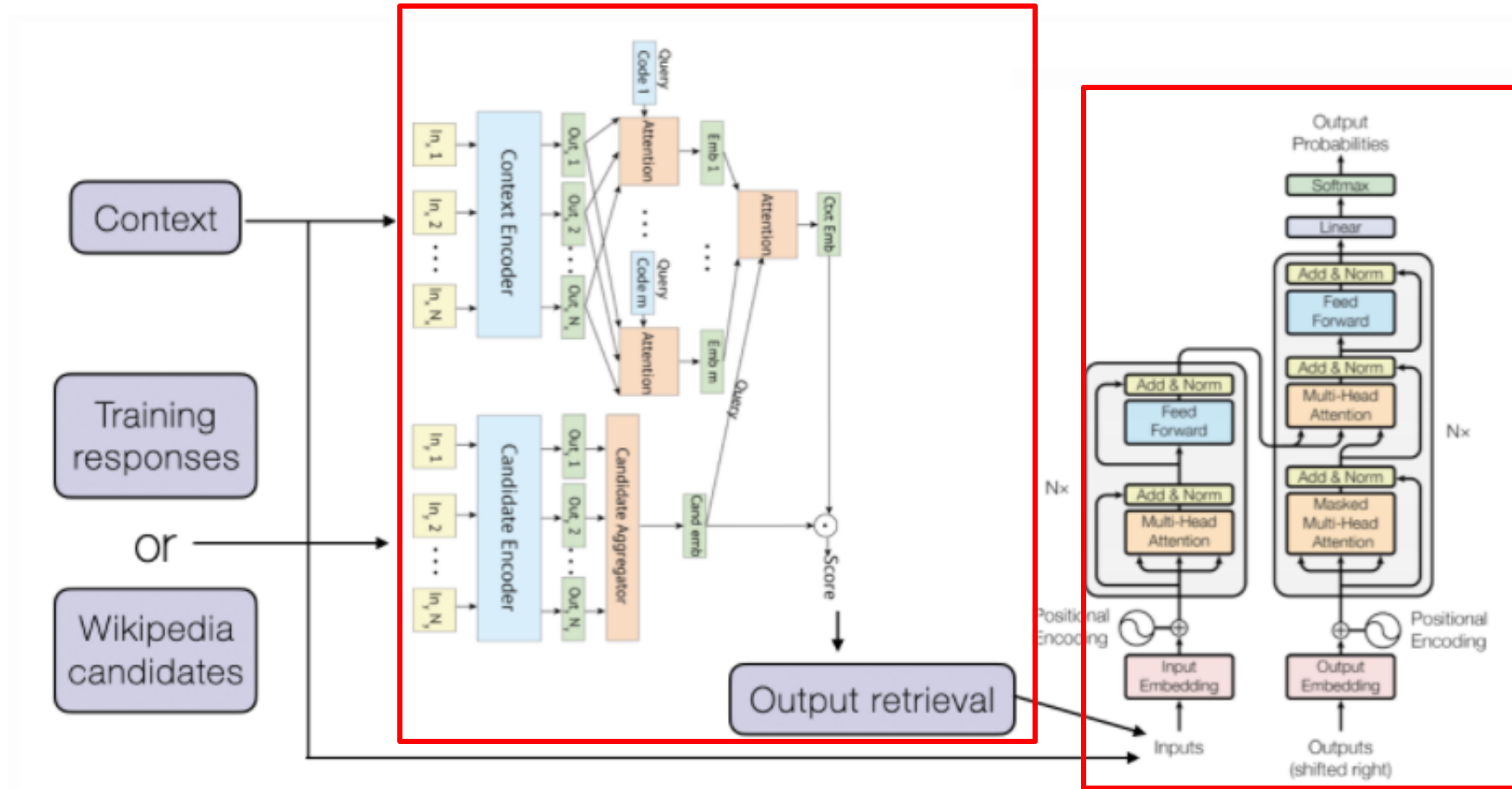
### Chatbot overview

- Requirements
  - providing engaging talking points
  - displaying knowledge, empathy and personality
  - maintaining consistent persona
- Key
  - training data: Requirements가 반영된 dataset 필요
  - generation strategy: 각 evaluation 방법에 맞는 generation strategy 다름  
e.g. human evaluation에서 utterance의 길이가 중요. 짧으면 낮은 점수.
- Limitation
  - lack of in-depth knowledge
  - stick to simpler language
  - repeat often used phrases

## Architecture

### Architecture

- Retrieve -> Generate



### Architecture

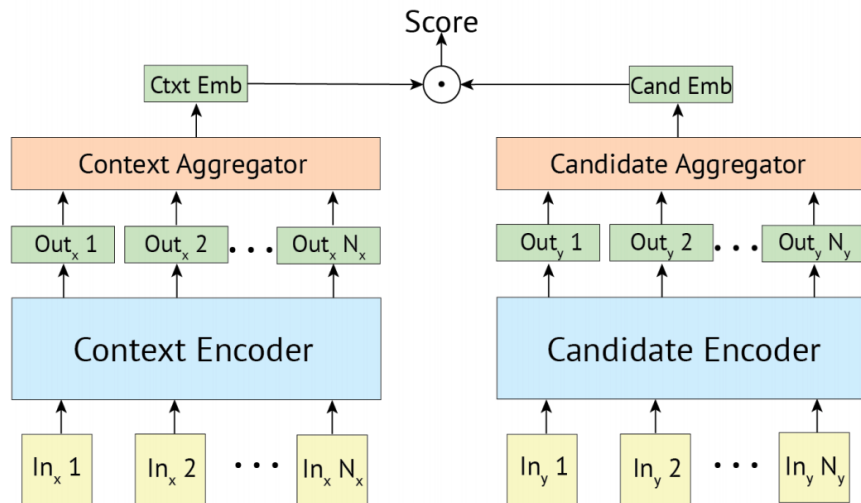
- 1) Retriever: DB에서 알맞은 response 검색
  - retrieval based model weakness: dependent on pre-constructed DB
- 2) Generator: response 생성
  - generative model weakness: repetitive response, lack of knowledge
- 3) Retrieve and Refine: Retrieval + Generative model
  - generation step 전에 retrieved response를 decoder에 input

## Retrieval

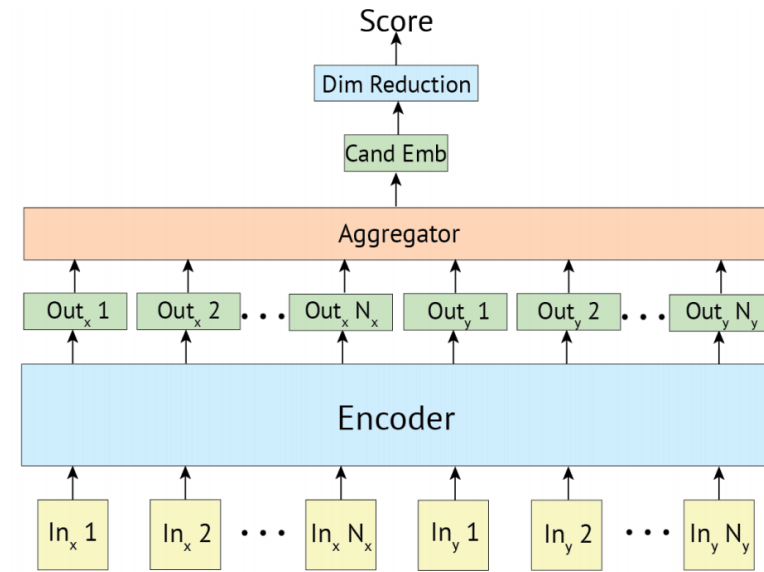
- Poly encoder: Bi encoder와 cross encoder의 장점을 사용

(Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring, ICLR 2020)

- Bi encoder: scoring할 때, 이미 **임베딩 완료된 candidate**와 scoring
  - > 빠르지만 query에 indenpendent
- Cross encoder: scoring할 때, input query 반영하여 candidate scoring
  - > query에 dependent라서 accurate but slower



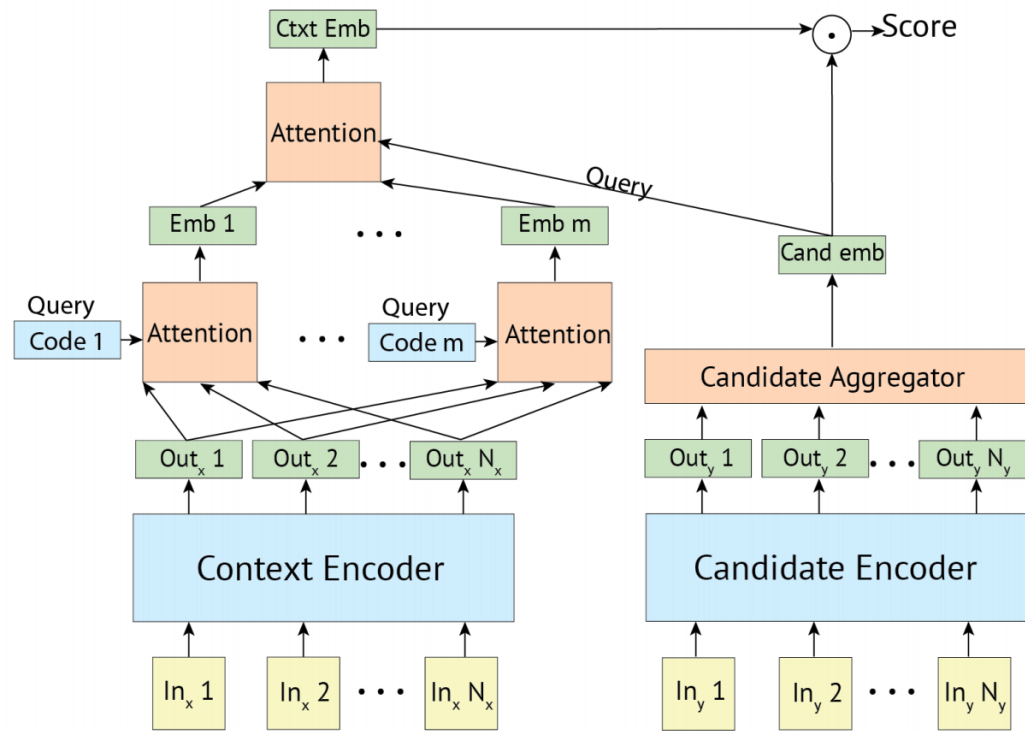
(a) Bi-encoder



(b) Cross-encoder

## Retrieval

- Poly encoder: Bi encoder와 cross encoder의 장점을 사용
  - cacheing candidate + jointly embedding input query



(c) Poly-encoder

### Generator

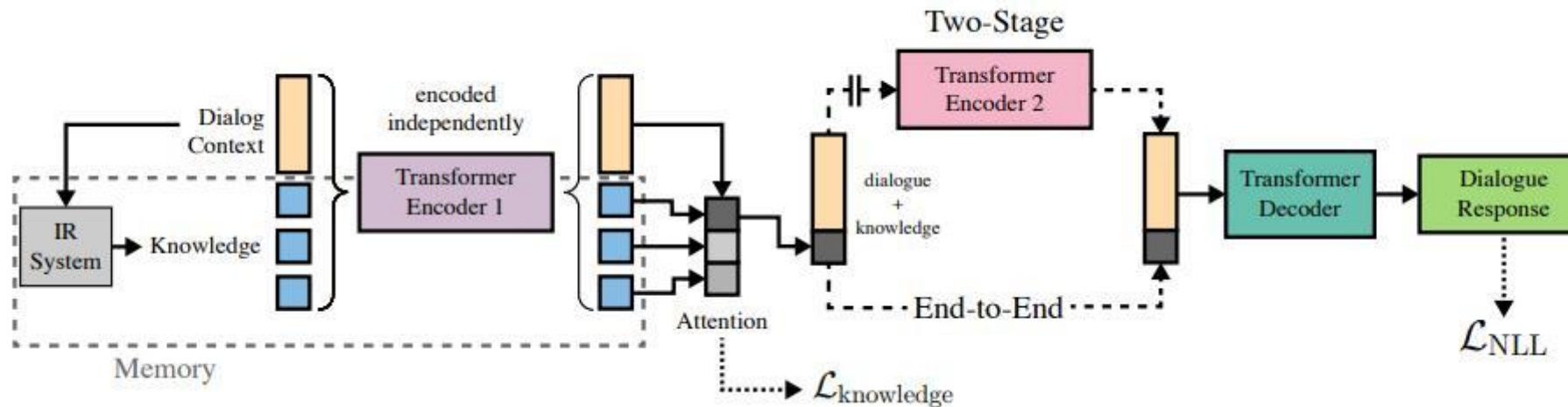
- seq2seq transformer
- "Towards a human-like open-domain chatbot" (a.k.a Meena) 와 구조 비슷
- Meena vs This work

sationalist. We use a seq2seq model (Sutskever et al., 2014; Bahdanau et al., 2015) with the Evolved Transformer (So et al., 2019) as the main architecture. The model is trained on multi-turn conversations where the input sequence is all turns of the context (up to 7) and the output sequence is the response. Our best model has 2.6B parameters and achieves a test perplexity of 10.2 based on a vocabulary of 8K BPE subwords (Sennrich et al., 2016).

tion heads. Our 2.7B parameter model roughly mimics the architectural choices of Adiwardana et al. (2020), with 2 encoder layers, 24 decoder layers, 2560 dimensional embeddings, and 32 attention heads.

Retrieve and Refine: combine retrieval step before generation

- Dialogue retrieval: poly encoder
- Knowledge retrieval: do\_require\_knowledge에 대해 binary clf -> retrieve knowledge if True
  - cite: "Wizard of wikipedia: Knowledge-powered conversational agents"
  - Condition the generation on the retrieved knowledge
  - Knowledgeable discussion 가능





### Retrieval model

- score gold response vs negative sample

### Generative model

- MLE

Given a dataset  $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}$ , minimize:

$$\mathcal{L}_{\text{MLE}}^{(i)}(p_{\theta}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = - \sum_{t=1}^{|\mathbf{y}^{(i)}|} \log p_{\theta}(y_t^{(i)} | \mathbf{x}^{(i)}, y_{<t}^{(i)}),$$

### Retrieve and refine

- Dialogue retrieve: generation할 때 retrieved response 반영하도록 훈련 (MLE사용)
  - issue: ignore the retrieval utterance
    - context와 retrieved response를 단순 concat하여 decoding하면 retrieved response는 무시됨
  - solution:  $\text{concat}(\text{context}, [\text{sep}], \text{retrieved}) \leftrightarrow \text{concat}(\text{context}, [\text{sep}], \text{gold response})$ 
    - 훈련시 alpha %만큼은 retrieved대신 gold response를 concat.
    - [sep] 이후 data 사용하도록 유도
- Knowledge retrieve: MLE

### 추가적인 generation loss function

- Unlikelihood training for generation
  - 목적: repetitive token과 over-presented token 생성문제 해결
  - MLE가 gold response 나올 확률을 maximize하는 것이라면
  - UL은 특정 n-gram이 나오면 penalty부여하여 나올 확률을 줄임
  - C\_t는 negative candidate (over-presented token)
    - negative candidate 선택: 실제 dialogue를 분석하여 predefined threshold 빈도 넘으면 C\_t에 편입

$$\mathcal{L}_{\text{ULE}}^{(i)} = \mathcal{L}_{\text{MLE}}^{(i)} + \alpha \mathcal{L}_{\text{UL}}^{(i)}$$

$$- \sum_{t=1}^{|y|} \sum_{y_c \in \mathcal{C}_t} \log(1 - p_{\theta}(y_c | \mathbf{x}, y_{<t}))$$

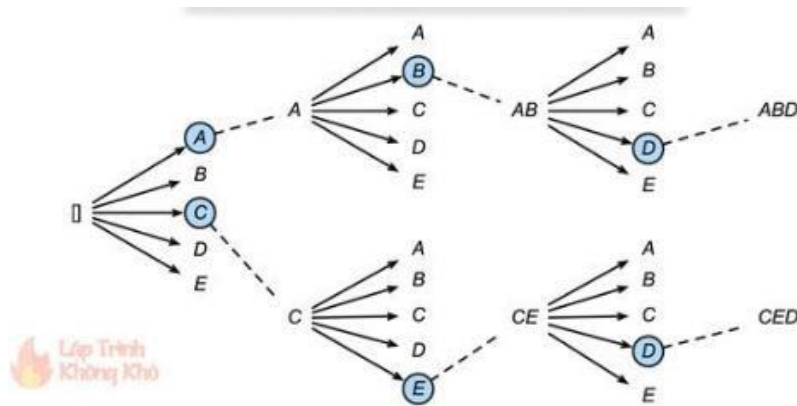
### 추가적인 generation loss function

- Unlikelihood training for generation
  - 목적: repetitive token과 over-presented token 생성문제 해결
  - MLE가 gold response 나올 확률을 maximize하는 것이라면
  - UL은 특정 n-gram이 나오면 penalty부여하여 나올 확률을 줄임
  - C\_t는 negative candidate (over-presented token)
    - negative candidate 선택: 실제 dialogue를 분석하여 predefined threshold 빈도 넘으면 C\_t에 편입

<i>n</i> -gram	MLE	UL	Human
Do you have	110	60	6
you have any	82	46	2
a lot of	74	46	14
What do you	57	20	6
you like to	54	43	1

Figure 5: Counts of 5 most common 3-grams from the BST Generative 2.7B model (MLE) from 100 conversation logs talking to crowdworkers, compared to those of the same model trained with unlikelihood (UL), and to human logs (for the same number of utterances).

- Beam search: Deterministic. Maintaining fixed-size set of partially decoded sequences.



- Sampling: Stochastic.
  - e.g. top-k sampling, ...
- Response length constraint
  - generative model은 짧은 문장을 선호.
  - 이를 해결하기 위해 classifier로 response의 length bin (<10, <20, <30, >30) 예측하도록 함
  - predicted length를 constraint로 사용
- Beam blocking
  - repetitive sequence generation문제 해결방법
  - 반복되는 n-gram은 generation 결과에서 filter함

## Training dataset

---

Pretrain: MLM objective

- pushshift.io Reddit: Reddit discussion

Fine-tune

- ConvAI2: crowdworker 간에 persona based conversation dataset. 서로의 persona 알아가는 conversation.
  - for engaging dialogue, persona consistency
- Empathetic Dialogue: emotional situation에서 crowdworker간 대화
  - for displaying empathy
- Wizard of Wikipedia: knowledgeable discussion
  - for displaying knowledge
- Blended skill dataset: engaging, knowledgeable, empathetic conversation

ACUTE-Eval: evaluation 질문에 대해 하나의 system 선택

- Engagingness question: "Who would you prefer to talk to for a long conversation?"
- Humanness question: "Which speaker sounds more human?"

Self-Chat ACUTE-Eval: 모델 결함 발견하기 쉬움 (speaker 1, 2를 하나의 system이 모두 play)

The screenshot displays a chat interface for the ACUTE-Eval evaluation. It shows a multi-turn conversation between two speakers, Speaker 1 and Speaker 2, represented by blue and purple bubbles respectively. The conversation starts with a greeting and continues with various topics including music, work, family, and hobbies. At the bottom, there is a question: "Who would you prefer to talk to for a long conversation?" with two radio button options: "I would prefer to talk to Speaker 1" and "I would prefer to talk to Speaker 2". Below this, there is a text input field for a justification: "Please provide a brief justification for your choice (a few words or a sentence)".

Hi how are you today

Hi, I'm pretty good! Just listening to some aerosmith, they're my fave :) whatre you up to?

I am listening to some italian music

Italian music, nice! What do you do for work?

I'm a professional wrestler

Oh wow! I'm a construction worker, I've been doing it since I was 16.

What do you do for a living

I'm a construction worker haha, just told you! Whats your favorite type of food?

My favorite is garlic bread

Yum! I love mexican food, so good :) do you have any hobbies?

I like to workout and go to the gym

We're a bit different- I love watching nascar and ufc. They're so fun!

Hello there, how are you?

I am doing great. How are you?

I am great, I did something crazy for me and colored my hair blue!

I have a daughter and a son who also love blue colored balls. You should meet them

Well that neat, I got a new car my mother gave so maybe I could see them!

It is a beautiful city. And, I try to be... Just cannot afford a bigger house atm.

I am sorry to hear that, I feel bad going out of town for spring break now.

Ok. I going to school in the spring for casino manager

Well I turn 29 next week, I wonder if that is a good age to apply as one.

My grandmother just died from lung cancer, sucks

Who would you prefer to talk to for a long conversation?

☐ I would prefer to talk to Speaker 1 ☐ I would prefer to talk to Speaker 2

Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

Figure A.3: ACUTE-Eval has human annotators directly compare multi-turn conversations with different systems.

Failure cases: repetition, forgetfulness, contradiction

- 같은 말 반복, 앞에 나온 정보 잊고 또 물어봄, 자신이 했던 말과 충돌하는 발언 => 대화를 이해하지 못함



Figure 6: **Examples of issues when talking to crowd-workers** with our Generative BST 2.7B model: non-trivial repetition (top example), forgetfulness (second example), contradiction (third example, Georgia is not in the Midwest).

Analysis

---

End