Recipes for Building an Open-Domain Chatbot

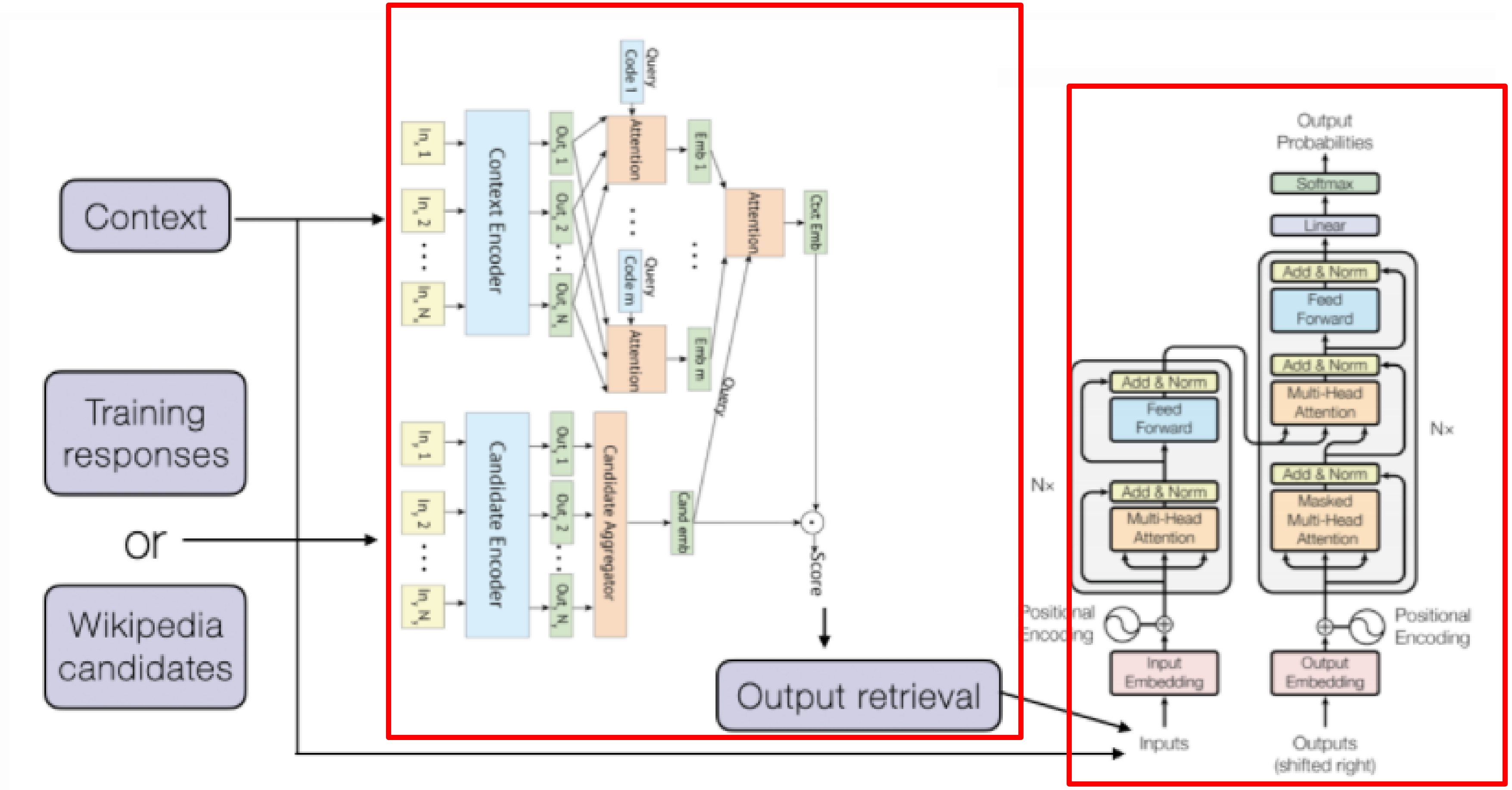
백형렬

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.

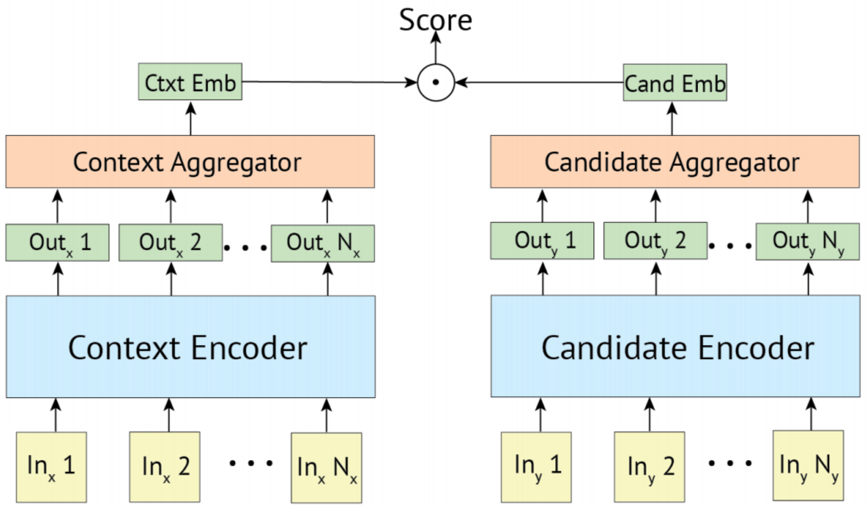
- Limitation
    - lack of in-depth knowledge
    - stick to simpler language
    - repeat often used phrases

Architecture
- Retrieve -> Generate

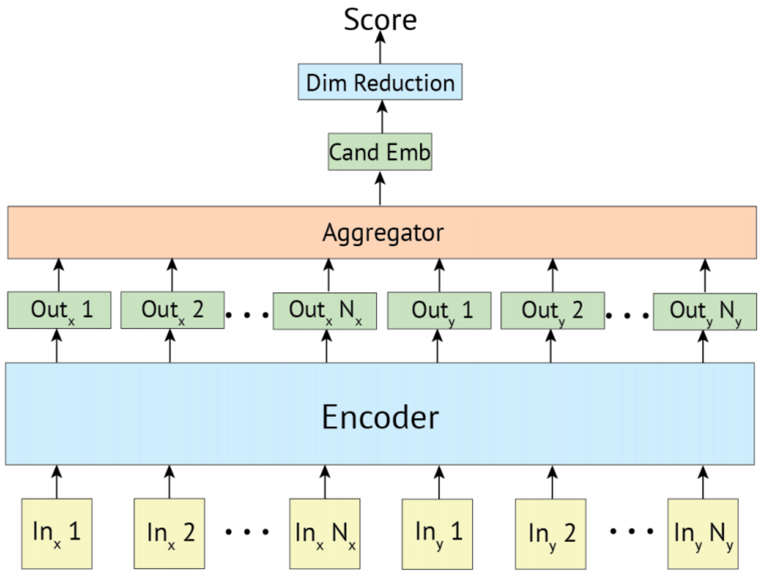


Architecture
1) Retriever: DB에서 알맞은 response 검색
- retrival based model weakness: dependent on pre-constructed DB
2) Generator: response 생성
- generative model weakness: repeatitive response, lack of knowledge
3) Retrieve and Refine: Retrieval + Generative model
- generation step 전에 retrieved response를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

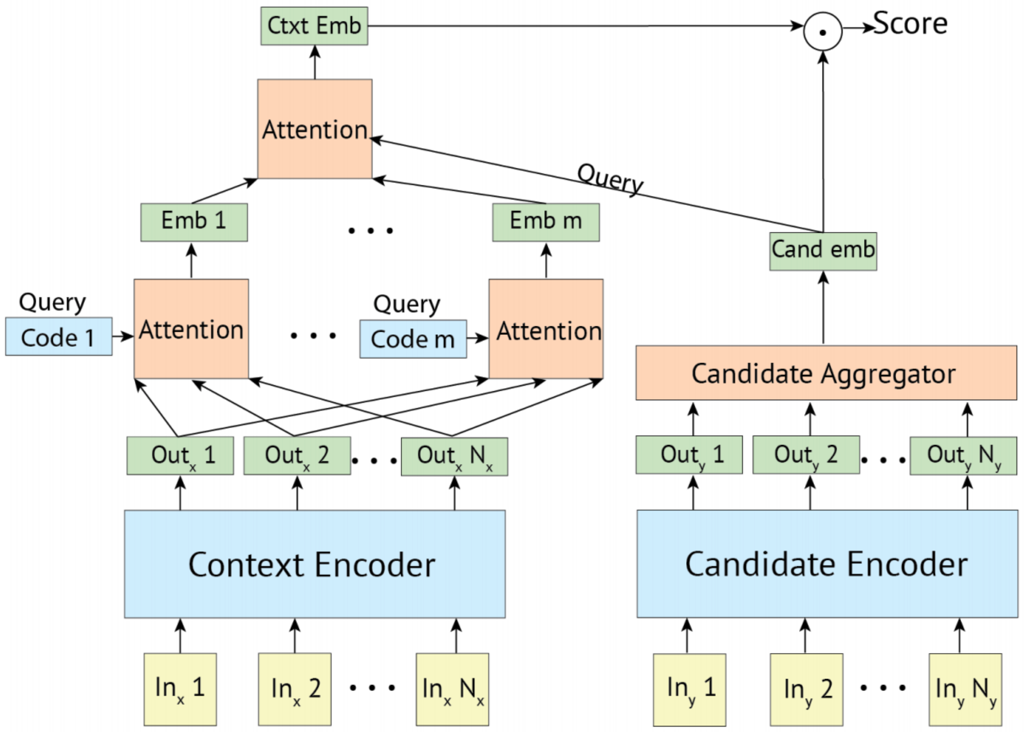


(a) Bi-encoder



(b) Cross-encoder

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



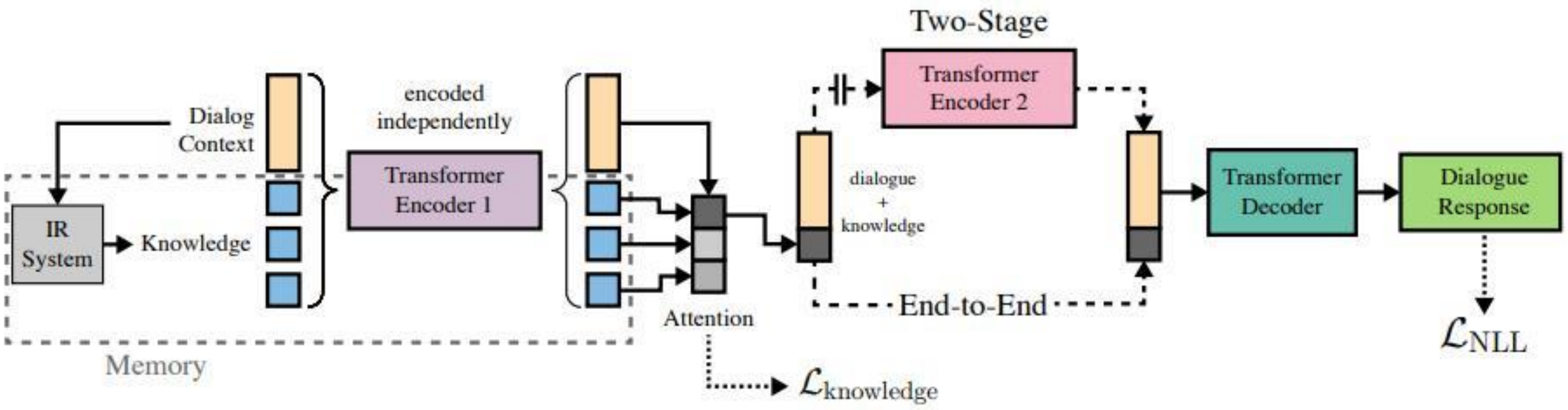
(c) Poly-encoder

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



Training objective - encoding

Retrieval model
- score gold response vs negative sample

Generative model
- MLE

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

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

Retrieve and refine
- Dialogue retrieve: generation할 때 retrieved response반영하도록 훈련 (MLE사용)
    - issue: ignore the retrieval utterance
        - context와 retrieved respsonse를 단순 concat하여 decoding하면 retrieved respsonse는 무시됨
    - solution: concat(context, [sep], retrieved) <-> concat(context, [sep], 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

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

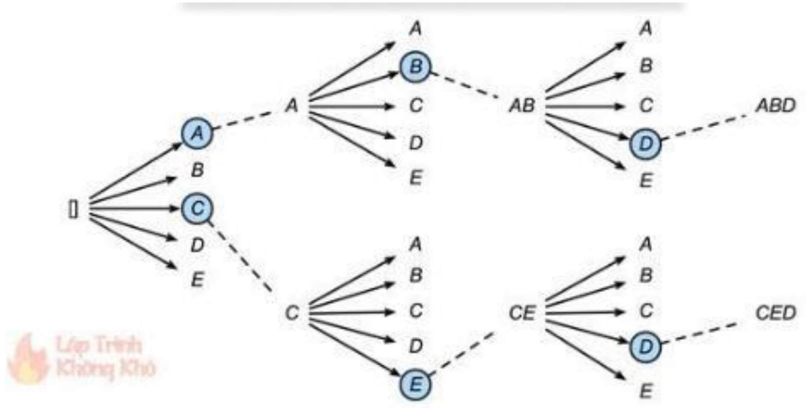
$$-\sum_{t=1}^{|y|} \sum_{y_c \in C_t} \log\left(1 - p_\theta(y_c | \mathbf{x}, y_{<t})\right)$$

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

n-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
  - repetitve sequence generation문제 해결방법
  - 반복되는 n-gram은 generation 결과에서 filter함

Pretrain: MLM objective
- pushshift.io Reddit: Reddit discussion

Fine-tune
- ConvAI2: crowdworker 간에 persona based coversation dataset. 서로의 persona 알아가는 conversation.
    - for engaging dialogue, persona consistency
- Empethetic Dialogue: emotional situation에서 crowdworker간 대화
    - for displaying empathy
- Wizard of Wikipedia: knowledgeable discussion
    - for dispalying knowledge
- Blended skill dataset: engaging, knowledeable, 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)

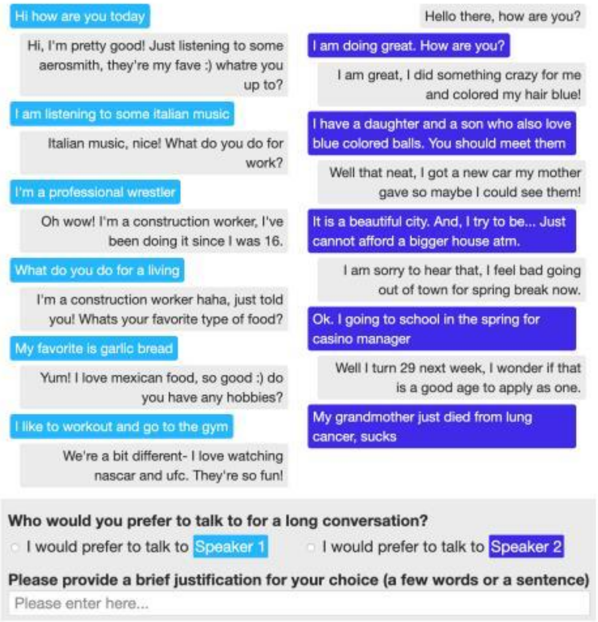


Figure A.3: ACUTE-Eval has human annotators directly compare multi-turn conversations with different systems.

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 crowdworkers** with our Generative BST 2.7B model: nontrivial repetition (top example), forgetfulness (second example), contradiction (third example, Georgia is not in the Midwest).

End