
Recipes for building an
open-domain chatbot
집현전 중급반 김대규
github.com/jiphyeonjeon

Introduction

Towards a human-like open-domain chatbot(Google)

- https://arxiv.org/abs/2001.09977
- The Meena model has 2.6 billion parameters and is trained on 341 GB of text, filtered from public domain social media conversations
  - Compared to an existing state-of-the-art generative model, OpenAI GPT-2, Meena has 1.7x greater model capacity and was trained on 8.5x more data.

https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

Introduction

BlenderBot: Recipes for building an open-domain chatbot (FAIR)

https://arxiv.org/abs/2001.09977

> While prior work has shown that scaling neural models in the number of parameters and the size of the data they are trained on gives improved results, we show that other ingredients are important for a high-performing chatbot.

- Blender: 지난 2년동안 Facebook AI Research(FAIR)에서 발표한 많은 작업을 융합한 것
- 사람이 보기에 좋은(사람의 평가에서 성능이 좋은) 오픈도메인 챗봇 태스크
  - Blended Skill Talk(BST) 데이터: 페르소나, 참여, 공감, 지식
  - Retrieve and Refine Model:
  - Decoding Strategy

Training Data

Dataset	Trait	Paper
ConvAI2	personality, engaging	Personalizing dialogue agents: I have a dog, do you have pets too? https://arxiv.org/abs/1801.07243
Empathetic Dialogues (ED)	empathy	Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset https://arxiv.org/abs/1811.00207
Wizard of Wikipedia (WoW)	knowledge	Wizard of Wikipedia: Knowledge-Powered Conversational agents https://arxiv.org/abs/1811.01241
Blended Skill Talk(BST)	blending skills	Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills https://arxiv.org/abs/2004.08449

Training Data: ConvAI2

서로에 대해 모르는 두명이 서로를 알아가는 태스크

- personality, engaging140k쌍 utterance data
- 각자 자신만 볼 수 있는 Persona를 부여
- 서로에 대해 알아가는 질문 답변 진행

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Table 2: Example dialog from the PERSONA-CHAT dataset. Person 1 is given their own persona (top left) at the beginning of the chat, but does not know the persona of Person 2, and vice-versa. They have to get to know each other during the conversation.

Personalizing Dialogue Agents: I have a dog, do you have pets too? https://arxiv.org/pdf/1801.07243.pdf

Data: Empathetic Dialogues(ED)

- 한 스피커는 개인적인 상황을 설명
- 다른 스피커는 "listener" 역할을 하여 대화 중에 공감을 표시

Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.
Listener: I'm sorry to hear that. I wish I could help you figure it out

Label: Proud
Situation: Speaker felt this when...
"I finally got that promotion at work! I have tried so hard for so long to get it!"
Conversation:
Speaker: I finally got promoted today at work!
Listener: Congrats! That's great!
Speaker: Thank you! I've been trying to get it for a while now!
Listener: That is quite an accomplishment and you should be proud!

Figure 2: Two examples from EMPATHETICDIALOGUES training set. The first worker (the speaker) is given an emotion label and writes their own description of a situation when they've felt that way. Then, the speaker tells their story in a conversation with a second worker (the listener).

Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset https://arxiv.org/abs/1811.00207

Data: Wizard of Wikipedia (WoW)

- 한 명은 지식이 풍부한 전문가 (Wizard) 역할을하고 다른 한 명은 호기심 많은 학습자 (Apprentice) 역할
- 대화 IR(TF-IDF) 시스템이 학습자의 질문에 가장 관련있는 위키피디아 paragraph들을 검색(top-7)
- 전문가(Wizard)가 가장 적절한 paragraph를 선택 후 답변 생성
- 1,250개 topic에 대하여 194k개 utterance
- 지식을 학습하는 것이 아닌 지식에 대한 표현을 학습 하는 것이 목표

Model: Two-Stage Generative Transformer MemNet

Topic:	Shower (*unseen*)
Human:	So I've been looking for a better showerhead with more water pressure for my shower.
Model:	I love pressure-versed water at nearly constant temperature to a shower
Human:	That sounds pretty great! Love those rainshower showers!
Model:	the ancient greeks were the first people to have showers.
Human:	Well I knew they were good for something.
Model:	the original showers were neither indoor structures nor man made but were common.
Human:	Ah that sounds intimidating!
Model:	it is required manual transport of both fresh and waste water.

Figure 2: Selected conversations between humans and models.

Wizard of Wikipedia: Knowledge-Powered Conversational agents https://arxiv.org/pdf/1811.01241.pdf

Data: Blended Skill Talk(BST)

ConvAI2, ED, WoW 데이터 특성들을 균일하게 결합(combine)하기 위해 이전 세 가지 태스크를 섞는것을 목표

- 두 발화자(Unguided Speaker, Guided Speaker)에게 ConvAI와 같이 페르소나 제공
- Unguided Speaker가 먼저 발언
- Guided Speaker는 세 가지 개별적인 태스크에서 학습된 봇들에 의해 제안된 발화를 선택하여 대화 진행 (수정하거나 무시하는 것 자유)
- 14 turn 더 이상 수

Persona for Unguided Speaker:
My son plays on the local football team.
I design video games for a living.

Persona for Guided Speaker:
My eyes are green.
I wear glasses that are cateye.

Wizard of Wikipedia topic: Video game design
Previous utterances (shown to speakers):
U: What video games do you like to play?
G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget

Actual utterances:
U: Exactly! I think many people fail to notice how beautiful the art of video games can be. (ConvAI2)
(G selected the WoW suggestion: "Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics.")
G: Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. (WoW)
U: Video games are undervalued by many and too easily blamed for problems like obesity or violence in kids (WoW)
G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. (ED)
U: I use games to relax after a stressful day, the small escape is relaxing. (ConvAI2/ED)
(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")
G: I enjoy a good gaming session after a hard dav at work as well. (ConvAI2/ED)
U: What other hobbies does your son have?(ConvAI2)
G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. (ConvAI2)
U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. (ConvAI2)
G: I wish I could play football, But I wear this cateve glasses and they would break if I tried. (ConvAI2)
U: Sounds nice. Are they new or vintage? (ConvAI2)
G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. (ConvAI2)

Figure 3: Sample conversation from the Blended Skill Talk dataset, which blends three skills that previous datasets (ConvAI2, WoW, ED) have focused on. Individual utterances are annotated with the single-skill datasets they are reminiscent of. The conversation here has been seeded with two utterances from WoW. For details about the Guided and Unguided workers (U,G) set up, see Smith et al. (2020).

Recipes for building an open-domain chatbot https://arxiv.org/pdf/2004.13637.pdf

Model architectures

1. **Retrieval model**
   - Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring (FAIR)
   - https://arxiv.org/abs/1905.01969
2. **Generative model**
3. **Retrieve-and-Refine: 검색-생성 혼합 모델**
   - Retrieve and Refine: Improved Sequence Generation Models For Dialogue (FAIR)
     - https://arxiv.org/abs/1808.04776
   - Wizard of Wikipedia: Knowledge-Powered Conversational agents (FAIR)
     - https://arxiv.org/abs/1811.01241

All three use Transformers (Vaswani et al., 2017) as a base.

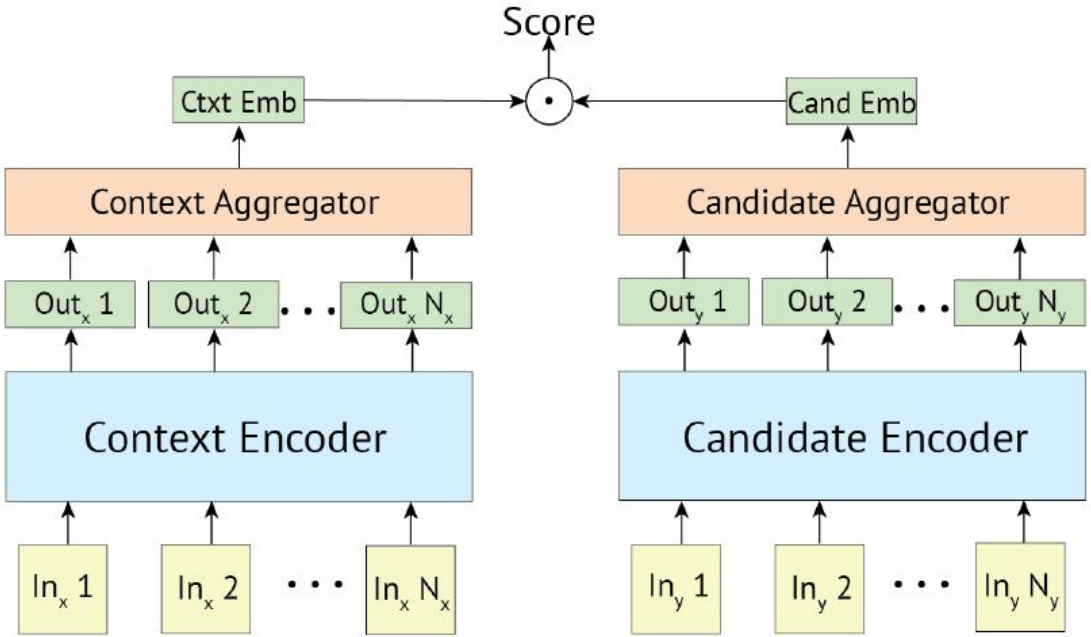
Retriever

> We employ the poly-encoder architecture of (Humeau et al., 2019). Poly-encoders encode global features of the context using multiple representations (n codes, where n is a hyperparameter), which are attended to by each possible candidate response

- Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)
- Reddit data pre-train
- 두가지 사이즈 고려: 256M, 622M parameters
- Bi-Encoder와 Cross-Encoder 장점을 혼합

Recipes for building an open-domain chatbot https://arxiv.org/pdf/2004.13637.pdf

Bi-Encoder



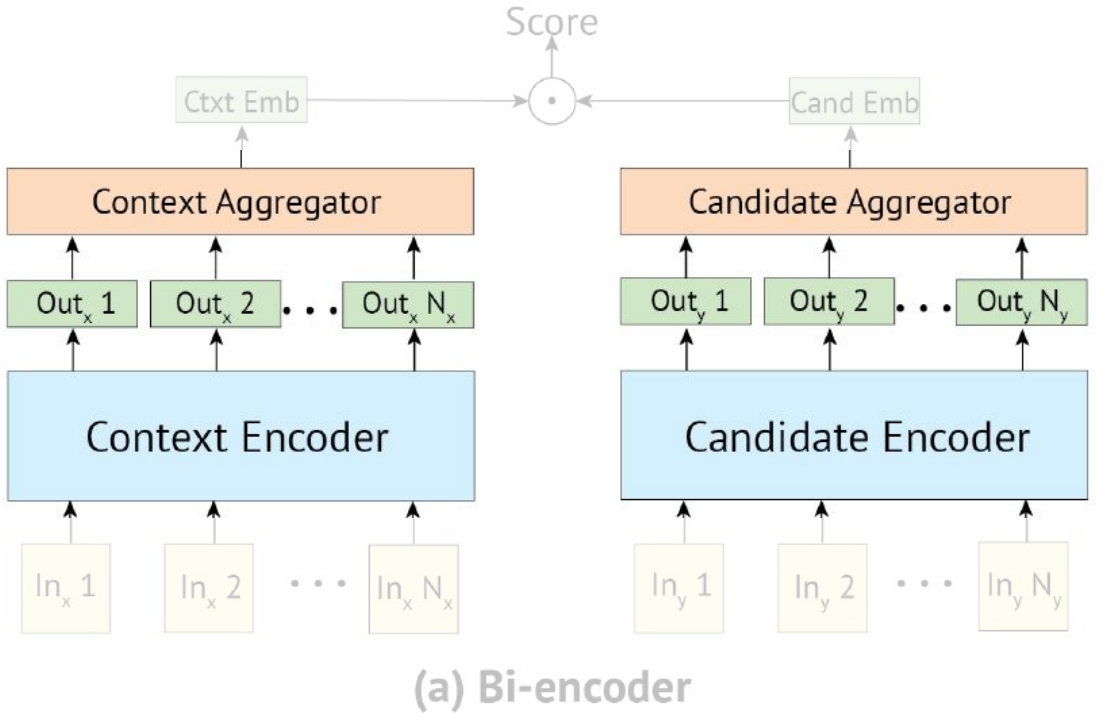
(a) Bi-encoder

Bi-Encoder

$$y_{ctxt} = red(T_1(ctxt))$$

$$y_{cand} = red(T_2(cand))$$

T = Transformer



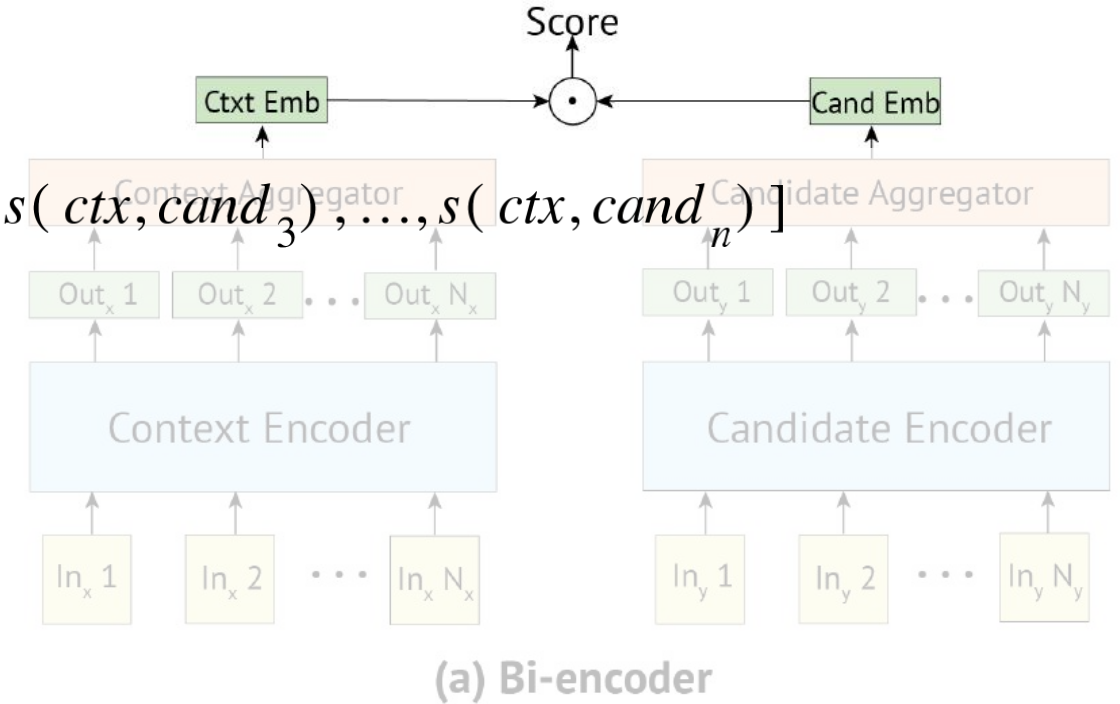
(a) Bi-encoder

Bi-Encoder

$$s(ctxt, cand_i) = y_{ctxt} \cdot y_{cand}$$

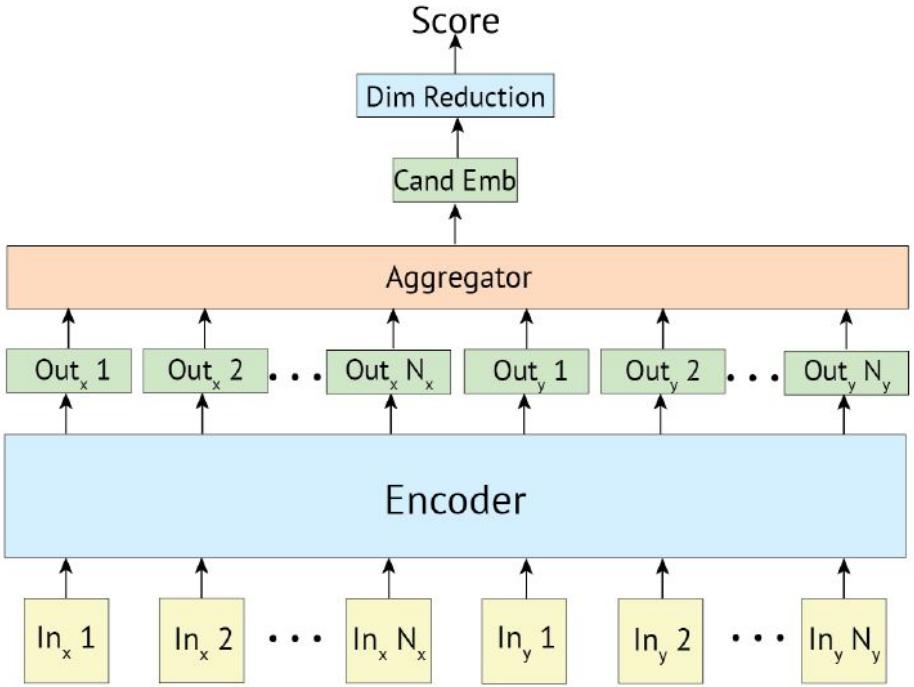
$$[s(ctx, cand_1), s(ctx, cand_2), s(ctx, cand_3), \ldots, s(ctx, cand_n)]$$

→ Cross Entropy



(a) Bi-encoder

Cross-Encoder

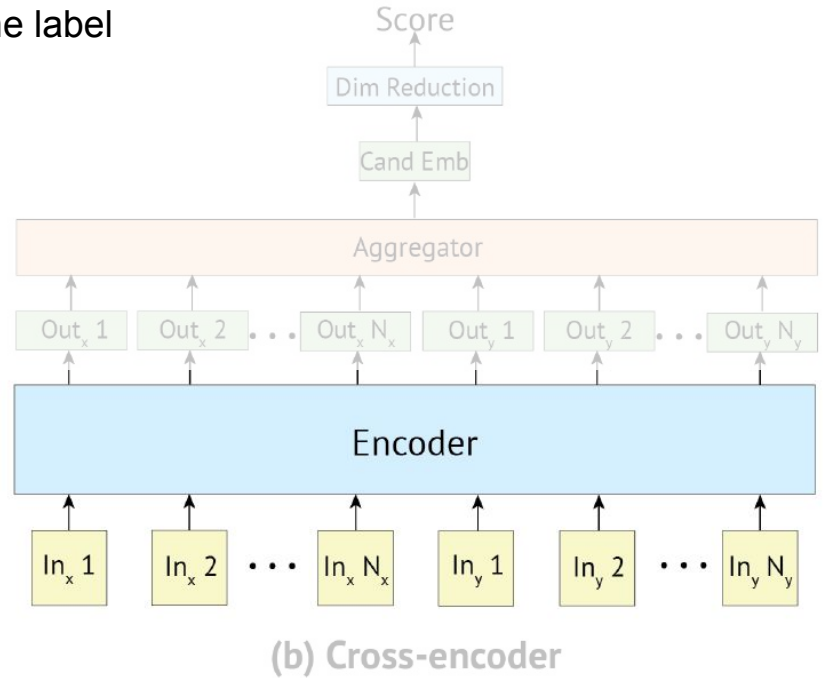


(b) Cross-encoder

Cross-Encoder

Full bi-directional attention between the input and the label

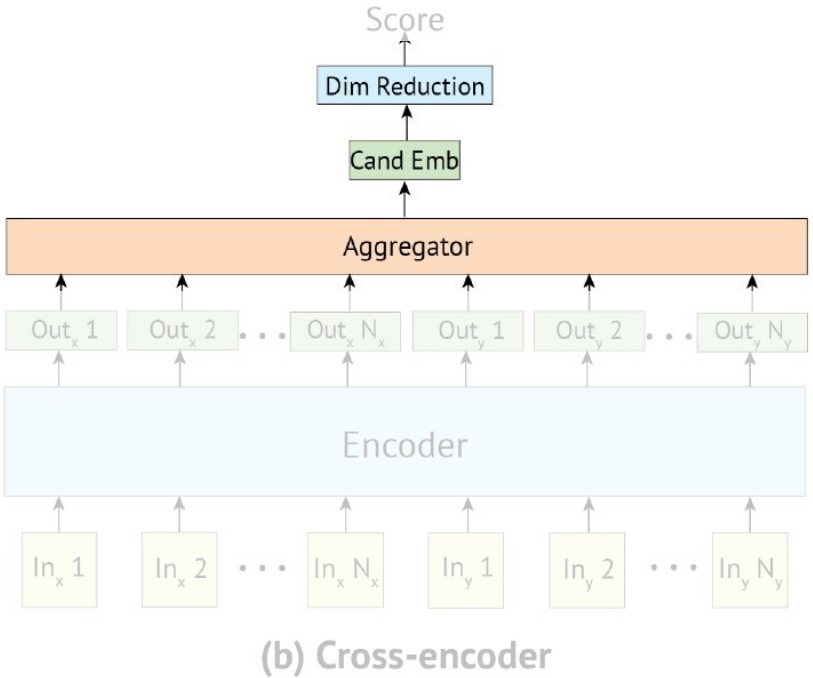
$$y_{ctxt,cand} = h1 = first(T(ctxt, cand))$$



(b) Cross-encoder

Cross-Encoder

$$s(ctxt, cand_i) = y_{ctxt, cand_i} \cdot W$$



(b) Cross-encoder

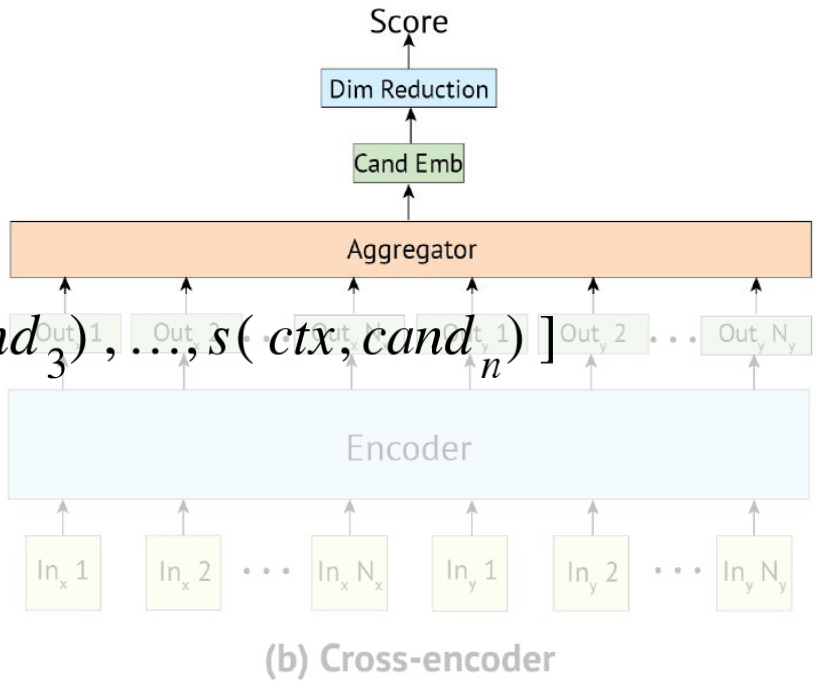
Cross-Encoder

$$s(ctxt, cand_i) = y_{ctxt, cand_i} \cdot W$$

$$[s(ctx, cand_1), s(ctx, cand_2), s(ctx, cand_3), \ldots, s(ctx, cand_n)]$$

→ Cross Entropy

where ‘cand1’ is the correct candidate and the others are negatives taken from the training set.



(b) Cross-encoder

Poly-Encoder

> The Poly-encoder architecture aims to get the best of both worlds from the Bi- and Cross-encoder. A given candidate label is represented by one vector as in the Bi-encoder, which allows for caching candidates for fast inference time, while the input context is jointly encoded with the candidate, as in the Cross-encoder, allowing the extraction of more information.

2 types Poly-Encoder model
- 256M: BERT-base: 12개 Encoder block, 12개 Attention head, 768 hidden size
- 622M: hyperparameter choices from those recommended in RoBERTa (Liu et al., 2019)

3가지의 attention을 수행
- Self-Attention over the Input Context's tokens
- Cross-Attention between codes(served as query) and outputs of the previous Self-Attention
- Cross-Attention between the 'm' global features of the Input Context and the embedding of the Candidate Label

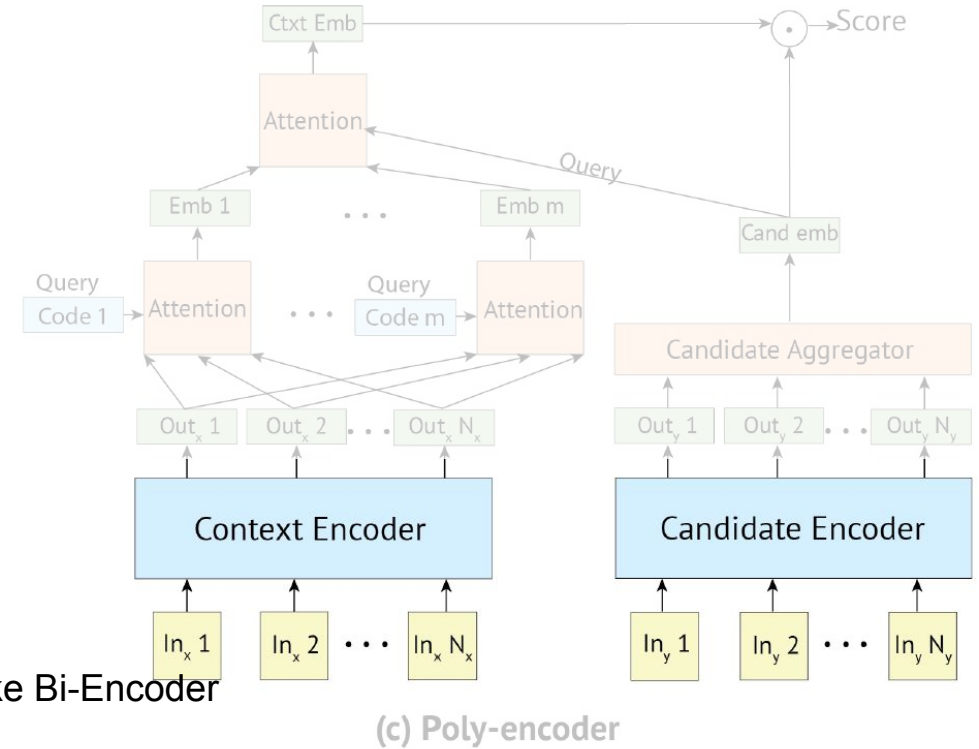
Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)

Poly-Encoder

$$y_{ctxt} = T_1(ctxt)$$

$$y_{cand} = red(T_2(cand))$$

Self-Attention over the Input Context's token like Bi-Encoder



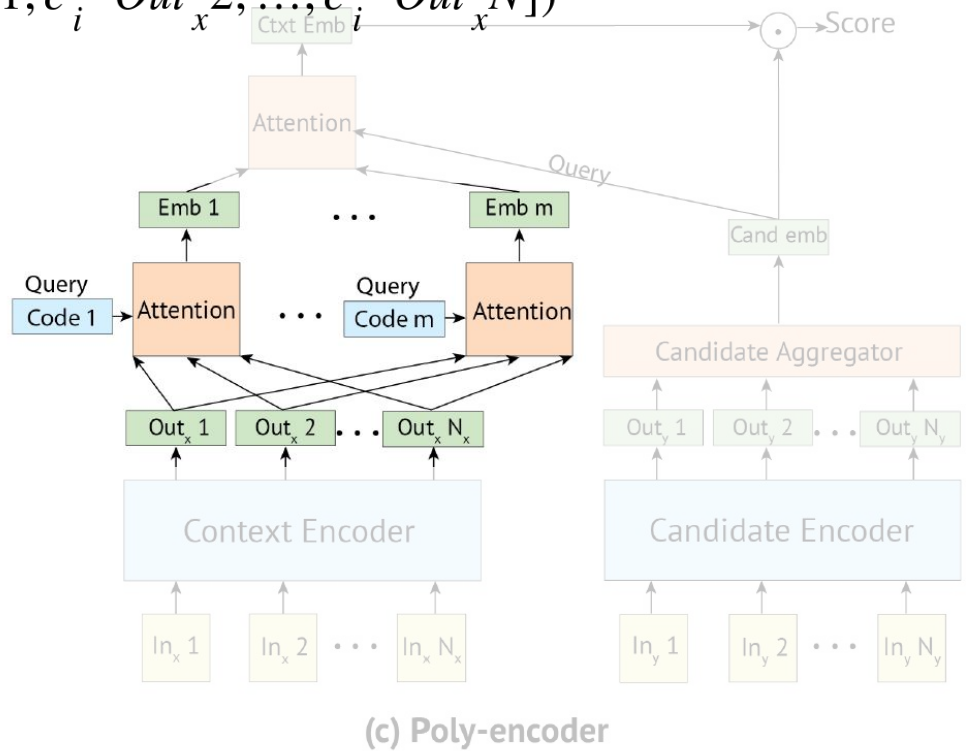
(c) Poly-encoder

Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)

Poly-Encoder

$$(w^{c_i}_1, w^{c_i}_2, \ldots, w^{c_i}_N) = softmax([c_i \cdot Out_x 1, c_i \cdot Out_x 2, \ldots, c_i \cdot Out_x N])$$

$$y^i_{ctxt} = \sum_i w^{c_i}_j \cdot Out_x j$$



(c) Poly-encoder

Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)

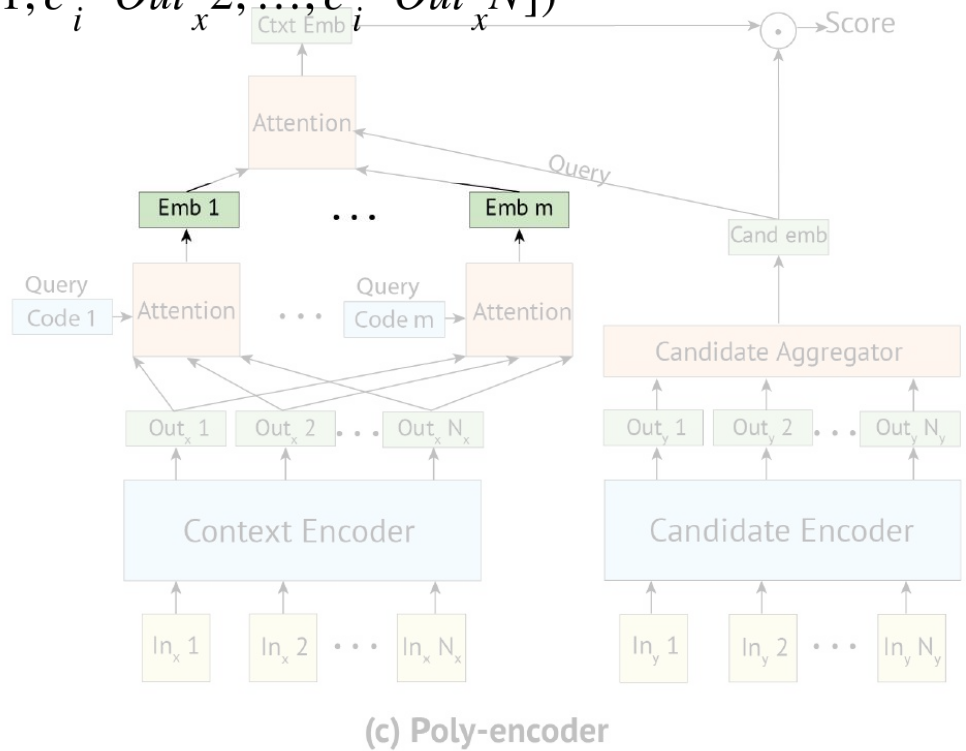
Poly-Encoder

$$(w_1^{c_i}, w_2^{c_i}, \ldots, w_N^{c_i}) = softmax([c_i \cdot Out_x 1, c_i \cdot Out_x 2, \ldots, c_i \cdot Out_x N])$$

$$y_{ctxt}^i = \sum_i w_j^{c_i} \cdot Out_x j$$

'm' global context features

$$[y_{ctxt}^1, y_{ctxt}^2, \ldots, y_{ctxt}^m]$$



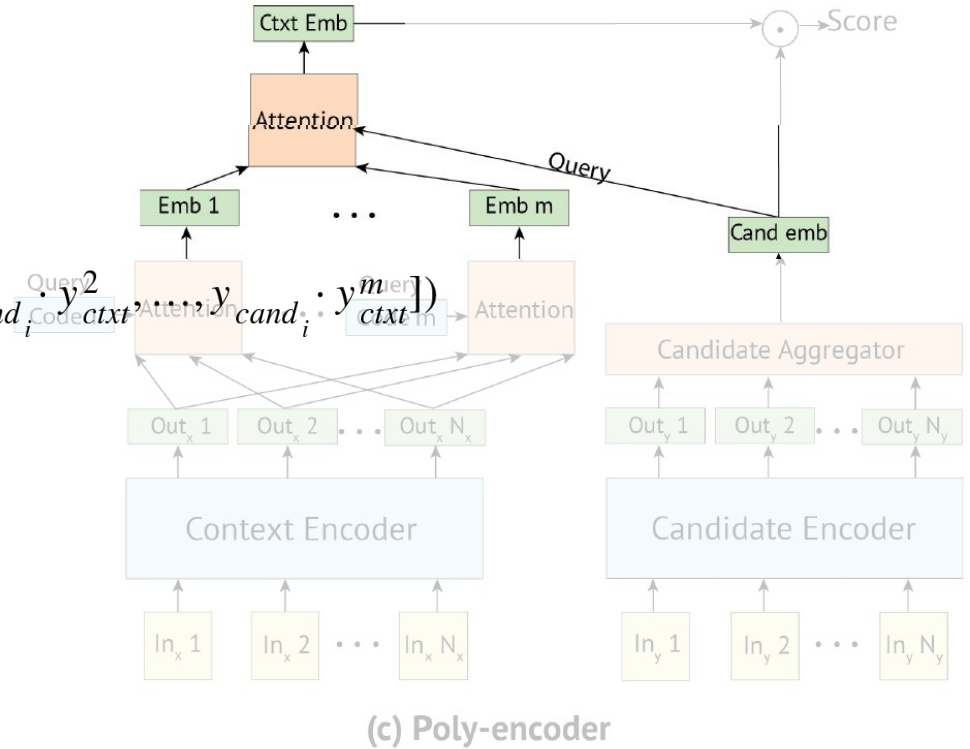
(c) Poly-encoder

Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)

Poly-Encoder

$$(w_1, w_2, \ldots, w_m) = softmax([y_{cand_i} \cdot y^1_{ctxt}, y_{cand_i} \cdot y^2_{ctxt}, \ldots, y_{cand_i} \cdot y^m_{ctxt}])$$

$$y_{ctxt} = \sum_i w_j \cdot y^i_{ctxt}$$



(c) Poly-encoder

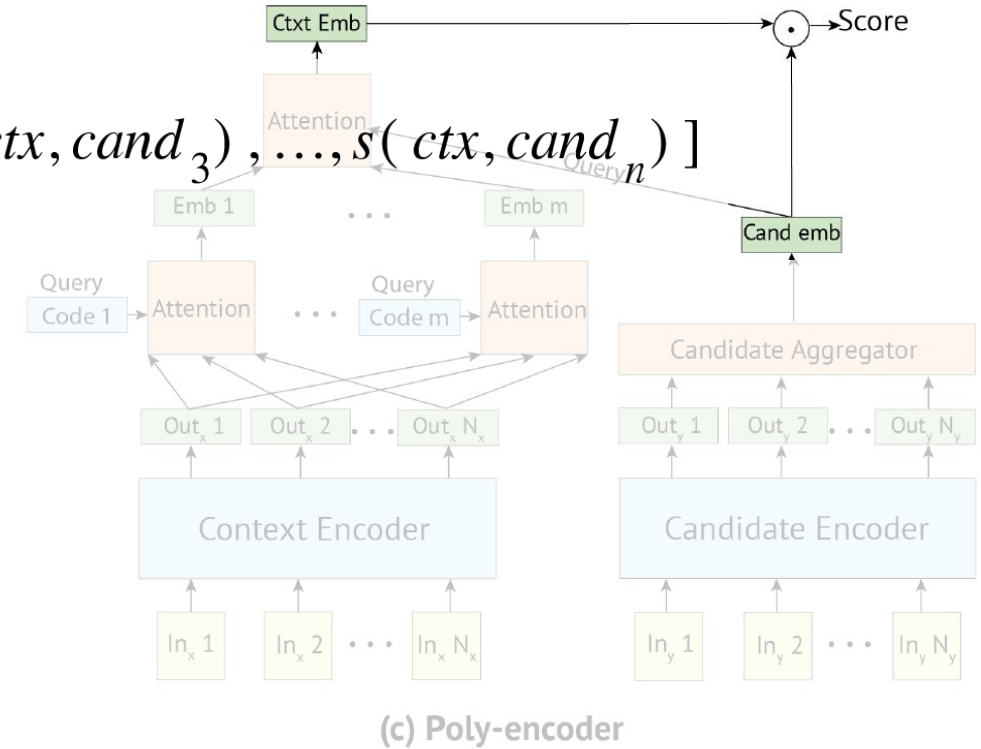
Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)

Poly-Encoder

$$s(ctxt, cand_i) = y_{ctxt} \cdot y_{cand_i}$$

$$[s(ctx, cand_1), s(ctx, cand_2), s(ctx, cand_3), \ldots, s(ctx, cand_n)]$$

→ Cross Entropy



(c) Poly-encoder

Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)

Bi-Encoder vs Cross-Encoder vs Poly-Encoder Experiment(performance)

Dataset	ConvAI2	DSTC 7		Ubuntu v2		Wikipedia IR
split	test	test		test		test
metric	R@1/20	R@1/100	MRR	R@1/10	MRR	R@1/10001
(Wolf et al., 2019)	80.7					
(Gu et al., 2018)	-	60.8	69.1	-	-	-
(Chen & Wang, 2019)	-	64.5	73.5	-	-	-
(Yoon et al., 2018)	-	-	-	65.2	-	-
(Dong & Huang, 2018)	-	-	-	75.9	84.8	-
(Wu et al., 2018)	-	-	-	-	-	56.8
pre-trained BERT weights from (Devlin et al., 2019) - Toronto Books + Wikipedia						
Bi-encoder	81.7 ± 0.2	66.8 ± 0.7	74.6 ± 0.5	80.6 ± 0.4	88.0 ± 0.3	-
Poly-encoder 16	83.2 ± 0.1	67.8 ± 0.3	75.1 ± 0.2	81.2 ± 0.2	88.3 ± 0.1	-
Poly-encoder 64	83.7 ± 0.2	67.0 ± 0.9	74.7 ± 0.6	81.3 ± 0.2	88.4 ± 0.1	-
Poly-encoder 360	83.7 ± 0.2	68.9 ± 0.4	76.2 ± 0.2	80.9 ± 0.0	88.1 ± 0.1	-
Cross-encoder	84.8 ± 0.3	67.4 ± 0.7	75.6 ± 0.4	82.8 ± 0.3	89.4 ± 0.2	-
Our pre-training on Toronto Books + Wikipedia						
Bi-encoder	82.0 ± 0.1	64.5 ± 0.5	72.6 ± 0.4	80.8 ± 0.5	88.2 ± 0.4	-
Poly-encoder 16	82.7 ± 0.1	65.3 ± 0.9	73.2 ± 0.7	83.4 ± 0.2	89.9 ± 0.1	-
Poly-encoder 64	83.3 ± 0.1	65.8 ± 0.7	73.5 ± 0.5	83.4 ± 0.1	89.9 ± 0.0	-
Poly-encoder 360	83.8 ± 0.1	65.8 ± 0.7	73.6 ± 0.6	83.7 ± 0.0	90.1 ± 0.0	-
Cross-encoder	84.9 ± 0.3	65.3 ± 1.0	73.8 ± 0.6	83.1 ± 0.7	89.7 ± 0.5	-
Our pre-training on Reddit						
Bi-encoder	84.8 ± 0.1	70.9 ± 0.5	78.1 ± 0.3	83.6 ± 0.7	90.1 ± 0.4	71.0
Poly-encoder 16	86.3 ± 0.3	71.6 ± 0.6	78.4 ± 0.4	86.0 ± 0.1	91.5 ± 0.1	71.5
Poly-encoder 64	86.5 ± 0.2	71.2 ± 0.8	78.2 ± 0.7	85.9 ± 0.1	91.5 ± 0.1	71.3
Poly-encoder 360	86.8 ± 0.1	71.4 ± 1.0	78.3 ± 0.7	85.9 ± 0.1	91.5 ± 0.0	**71.8**
Cross-encoder	**87.9 ± 0.2**	**71.7 ± 0.3**	**79.0 ± 0.2**	**86.5 ± 0.1**	**91.9 ± 0.0**	-

Table 4: Test performance of Bi-, Poly- and Cross-encoders on our selected tasks.

Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)

Bi-Encoder vs Cross-Encoder vs Poly-Encoder Experiment(time)

average time per example for each architecture

Dataset	ConvAI2	DSTC7	UbuntuV2
Bi-encoder	2.0	4.9	7.9
Poly-encoder 16	2.7	5.5	8.0
Poly-encoder 64	2.8	5.7	8.0
Cross-encoder64	9.4	13.5	39.9

Table 6: Training time in hours.

	Scoring time (ms)			
	CPU		GPU	
Candidates	1k	100k	1k	100k
Bi-encoder	115	160	19	22
Poly-encoder 16	122	678	18	38
Poly-encoder 64	126	692	23	46
Poly-encoder 360	160	837	57	88
Cross-encoder	21.7k	2.2M*	2.6k	266k*

Table 5: Average time in milliseconds to predict the next dialogue utterance from C possible candidates on ConvAI2. * are inferred.

Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring (https://arxiv.org/abs/1905.01969)

Generator

- Transformer Decoder(GPT-base) architecture
- Byte-Level BPE tokenization (Radford et al., 2019) trained on the pre-training data, as implemented in HuggingFace's Tokenizers
- 3가지 사이즈 고려
  - 90M
    - Shuster et al., 2019 따라 구현
    - 12 Decoder Layer, 768 Embedding dim, 12 attention heads
  - 2.7B
    - 최대 컨텍스트와 응답 길이가 128 BPE 토큰
    - 2개 Encoder Layer, 24개 Decoder Layer, 2560 Embedding dim, 32 attention heads
  - 9.4B
    - Adiwardana et al. (2020) 따라 구현
    - 최대 컨텍스트와 응답 길이가 128 BPE 토큰
    - 4개 Encoder Layer, 32개 Decoder Layer, 4096 Embedding dim, 32 attention heads

The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents https://arxiv.org/pdf/1911.03768.pdf
Towards a human-like open-domain chatbot https://arxiv.org/abs/2001.09977

Likelihood Training for Generation

Neural Text Generation with Unlikelihood Training(https://arxiv.org/abs/1908.04319)
standard Maximum Likelihood Estimation (MLE) approach

Given a dataset $D = (x(i), y(i))$, minimize:

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

https://medium.com/dair-ai/recipes-for-building-an-open-domain-chatbot-488e98f658a7

Generator Experiment

Name	Total Params	V	L_{enc}	L_{dec}	d	h	Steps	PPL
90M	87,508,992	55K	8	8	512	16	2.86M	25.6
2.7B	2,696,268,800	8K	2	24	2560	32	200K	13.3
9.4B	9,431,810,048	8K	4	32	4096	32	200K	12.2

Table 2: **Perplexity on the validation set of pushshift.io Reddit** for several generative Transformer models with given architecture settings. Note that perplexity is not directly comparable between the 90M models and the larger models as the 90M models use a different dictionary. Columns include the vocabulary size (V), number of encoder and decoder layers (L_{enc}, L_{dec}), embedding dimensionality (d), Multihead Attention Heads (h), and training steps.

Retrieve and Refine

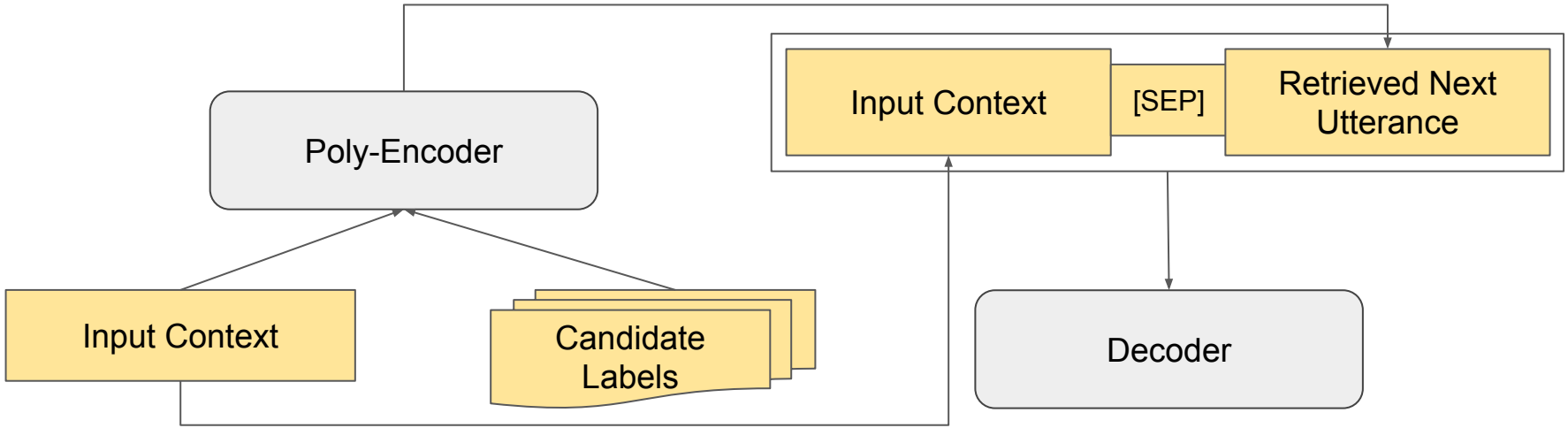
Generative Model

- models produce dull and repetitive responses
- models are hallucinate knowledge, and are unable to read and access external knowledge.

RetNRef Model

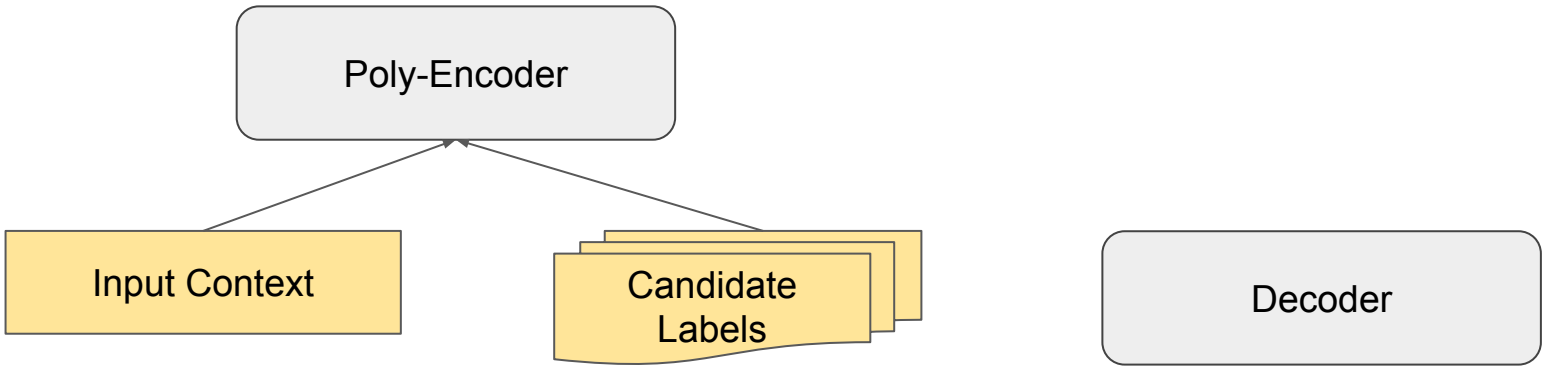
- Retrieve and Refine: Improved Sequence Generation Models For Dialogue (FAIR) (https://arxiv.org/abs/1808.04776)
- Retrieval before Generation
- 생성 모델이 검색 응답 후보(사람이 작성한)를 기반으로 사람다움을 언제 복제하는지 그리고 언제 안할지를 학습하여 생성모델 보다 향상된 응답 달성하는 것이 목표
- Dialogue Retrieval and Refine
- Knowledge Retrieval and Refine

Dialog Retrieve and Refine



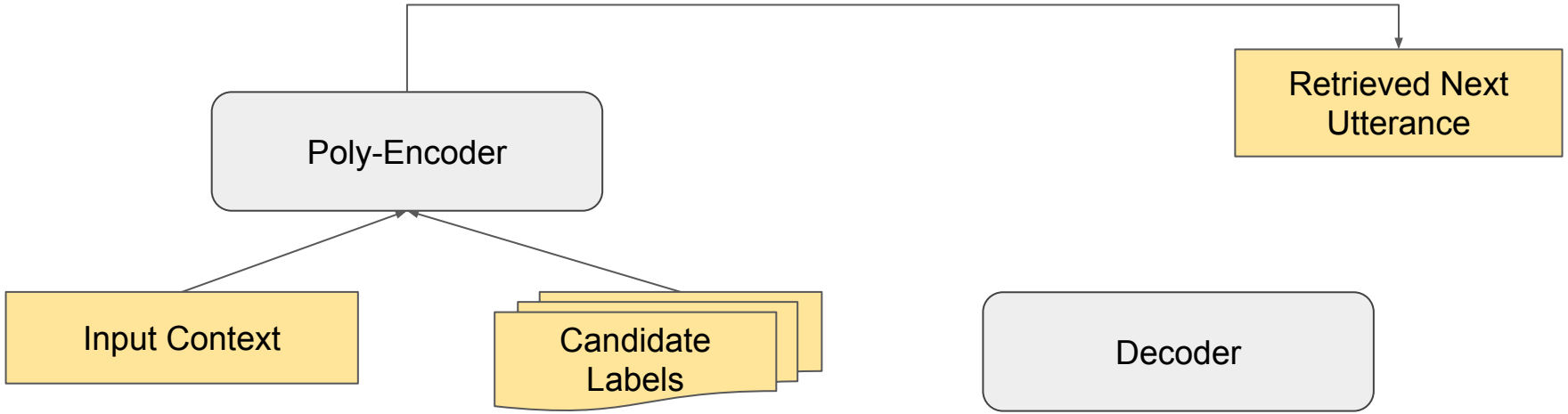
Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

Dialog Retrieve and Refine



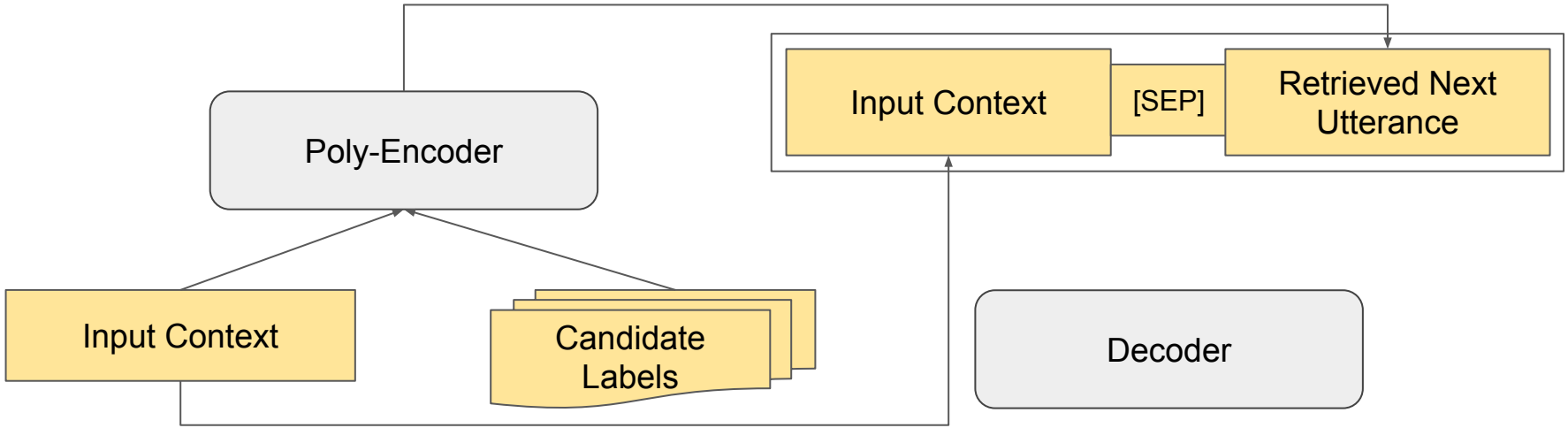
Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

Dialog Retrieve and Refine



Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

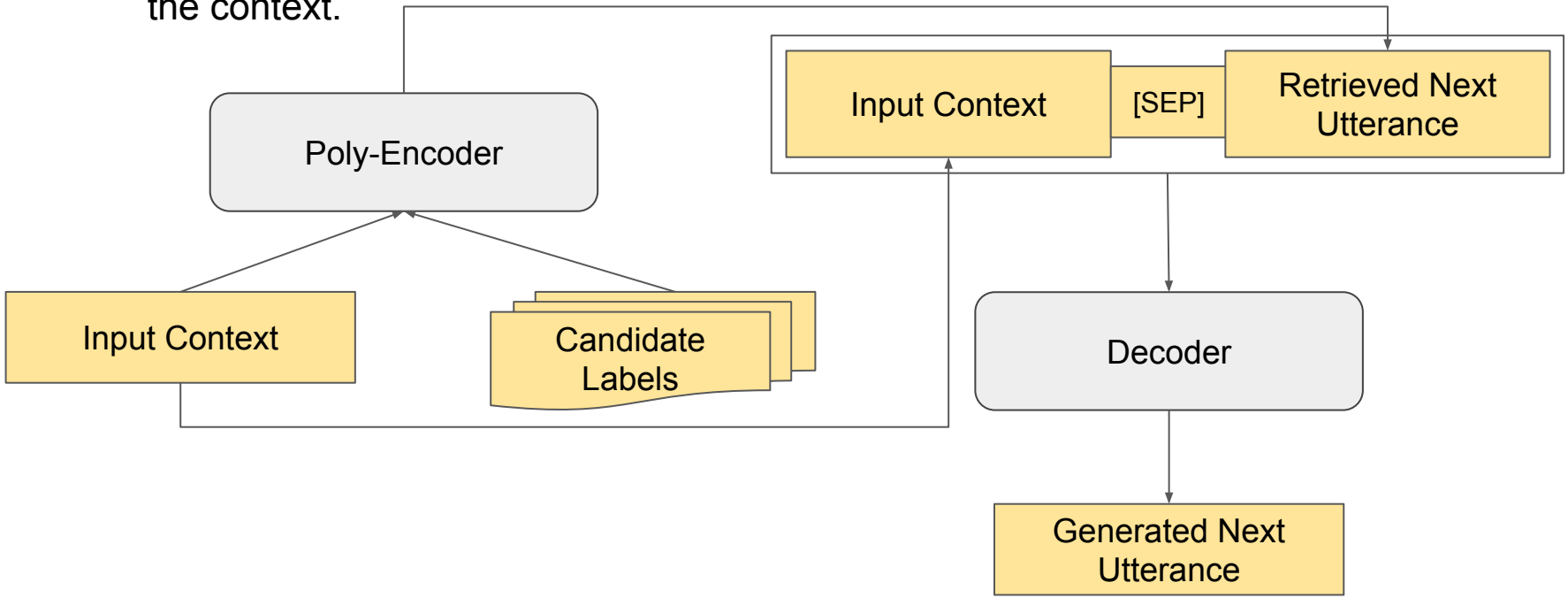
Dialog Retrieve and Refine



Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

Dialog Retrieve and Refine

Decoder learn when to simply use the “Retrieved Next Response” directly and when to ignore the retrieved response and generate one based only on the context.



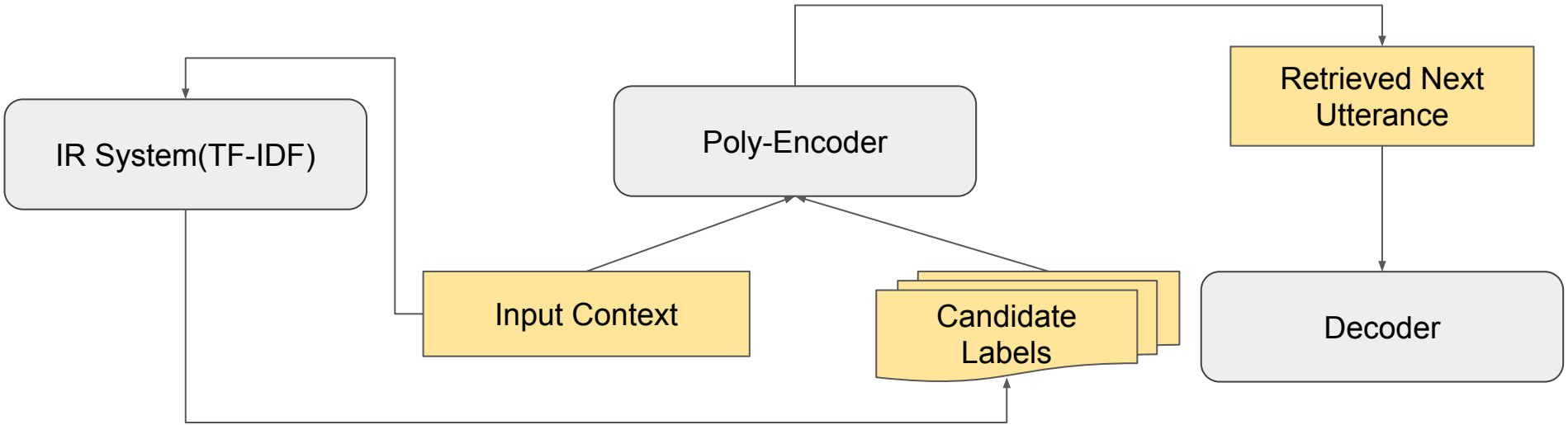
Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

α-blending for Retrieve and Refine

> As the correspondence between gold label and retrieved utterance is not necessarily clear, a trained model often opts to simply ignore the retrieval utterance .... To ensure it is used, one can replace the retrieved response instead with the gold response α% of the time

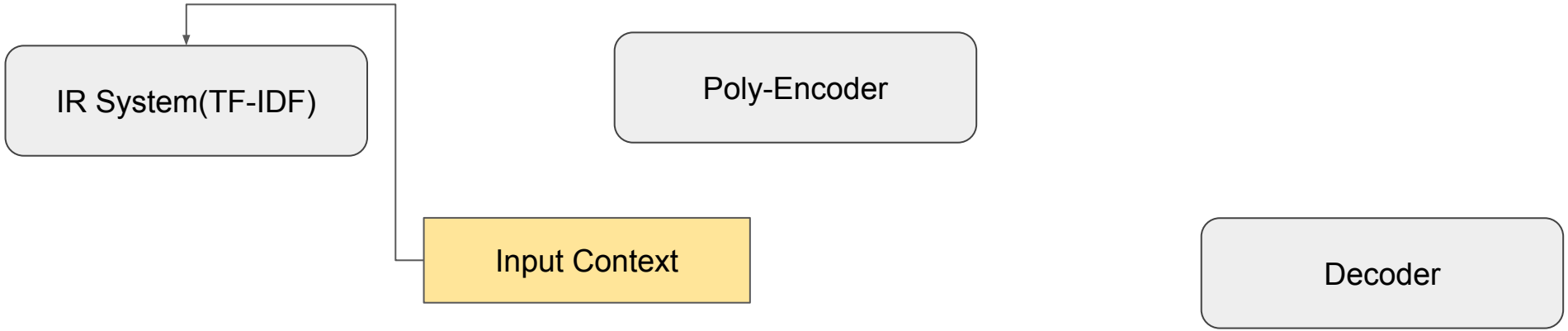
- gold label과 검색된 발화(retrieved utterance)간에 연관성은 반드시 명확하지 않기 때문에, Weston et al. (2018)에서 보여준바와 같이 학습된 모델은 자주 검색 발화(retrieval utterance)를 단순히 무시하는걸 선택합니다. (Retrieve and Refine: Improved Sequence Generation Models For Dialogue (https://arxiv.org/abs/1808.04776))
- α% time 만큼 gold response로 대체하는 것
- 검색과 생성 시스템 사이에 자연스러운 전환을 하게 합니다.

Knowledge Retrieve and Refine



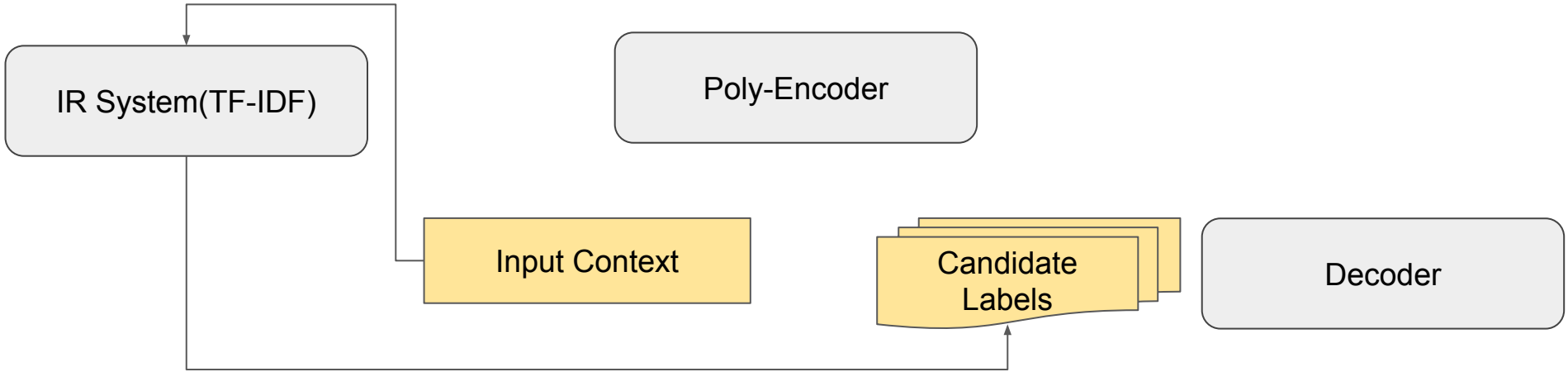
Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

Knowledge Retrieve and Refine



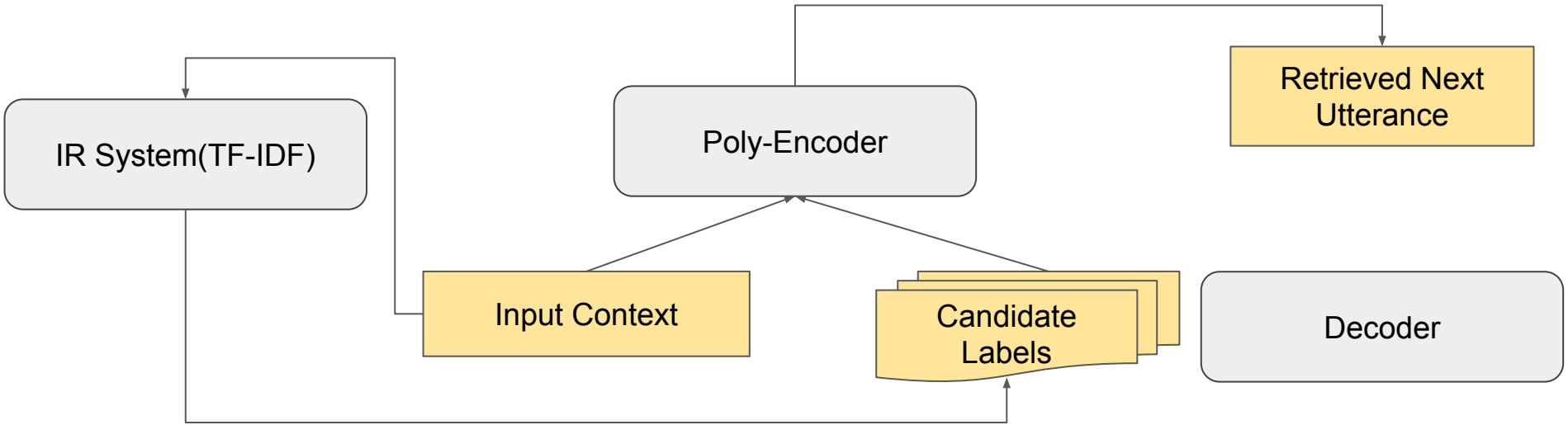
Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

Knowledge Retrieve and Refine



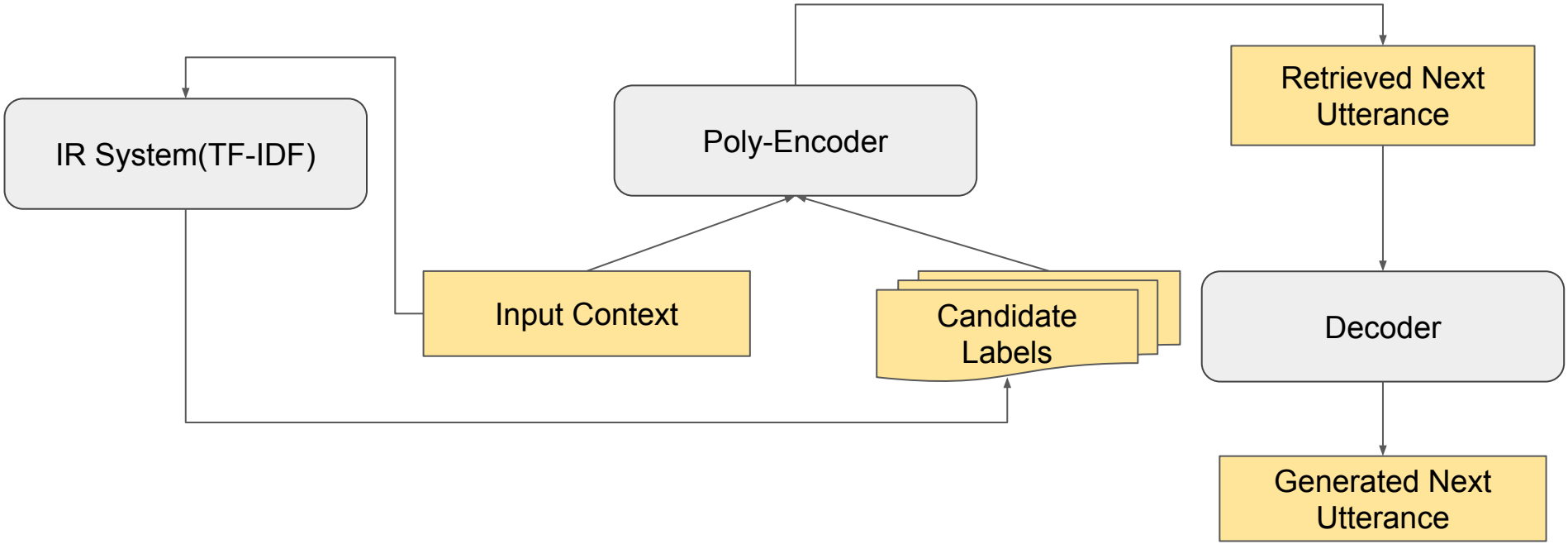
Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

Knowledge Retrieve and Refine



Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

Knowledge Retrieve and Refine



Blender Bot — Part 3: The Many Architectures https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6

Experiment: Generator vs. RetNRef - perplexity

Model	Size	ConvAI2	WoW	ED	BST	Avg.
pushshift.io Reddit Generative	90M	18.33	31.18	14.44	18.09	20.51
BST Generative	90M	11.36	17.56	11.48	14.65	13.76
BST RetNRef	256M/90M	11.79	18.37	11.87	14.62	14.16
pushshift.io Reddit Generative	2.7B	15.70	13.73	11.06	14.36	13.71
BST Generative	2.7B	8.74	8.78	8.32	10.08	8.98
BST RetNRef	622M/2.7B	9.31	9.28	9.93	10.59	9.78
pushshift.io Reddit Generative	9.4B	15.02	12.88	10.41	13.5	12.95
BST Generative	9.4B	8.36	8.61	7.81	9.57	8.59

Table 3: **Perplexity of the pre-trained and fine-tuned models on the validation set for BST datasets.** Note that perplexity is not directly comparable between the 90M models and the larger models as 90M models use a different dictionary. Fine-tuning gives gains for each skill (task) compared to pre-training on pushshift.io Reddit alone.

Experiment: Retrieval vs. Generator vs. RetNRef - ACUTE-Eval

Win %	Loss % Ret	Loss % Gen	Loss % RetNRef
Retrieval		29 *	30 *
Generative	71 *		44
RetNRef	70 *	56	

Figure 14: Human-bot ACUTE-Eval (engagingness): Retrieve and Refine($\alpha = 0.5$) and Generative (90M, beam search decoding, min beam size 20) beat Retrieval (256M). All results are significant ($p < 0.01$) except for RetNRef vs. Generative.

Evaluation Methods: ACUTE-Eval

ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons

- ACUTE-Eval 은 인간 주석자에게 직접적으로 다른 시스템의 멀티턴 대화를 비교하도록 합니다.
- Engagingness question: “Who would you prefer to talk to for a long conversation?”
  - 참여도 질문 : "오래 대화하기 위해 누구와 이야기하고 싶습니까?"
- Humanness question: “Which speaker sounds more human?”
  - 인간성 질문 :“어떤 스피커가 더 인간적으로 들리는가?”

Evaluation Methods: ACUTE-Eval

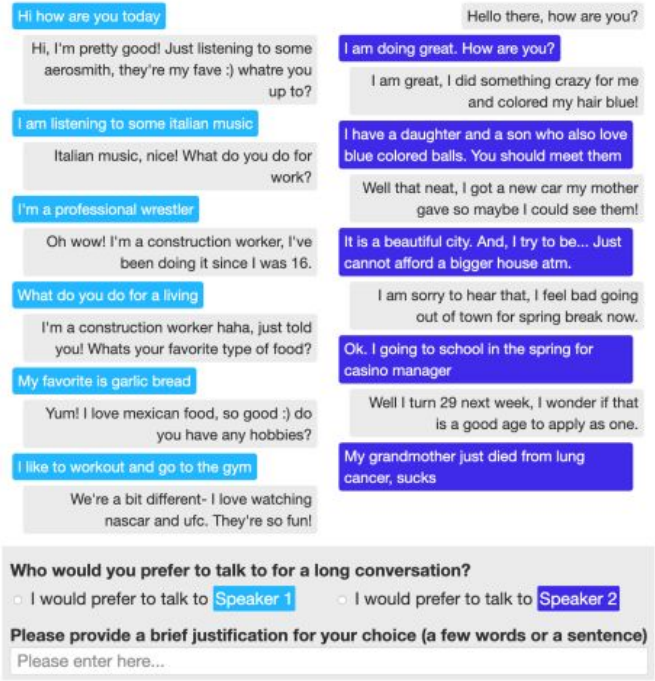


Figure 4: ACUTE-Eval has human annotators directly compare multi-turn conversations with different systems.

Unlikelihood training for generation

$$-\alpha \underbrace{-\sum_{c \in C_t} \log(1 - p_\theta(c|x_{<t}))}_{\text{unlikelihood}} \underbrace{- \log(p_\theta(x_t|x_{<t}))}_{\text{likelihood}}$$

- Neural Text Generation with Unlikelihood Training (https://arxiv.org/abs/1908.04319)
- 인간이 잘 사용하지 않는 토큰을 모델이 사용하는 것들에 대해 확률 분포를 낮춤으로써 적절한 횟수의 사용을 하도록 하는게 목표
- unlikelihood는 negative candidate C_t에 속하면 확률 분포를 낮춘다
- gold response에서 계산된 단어 카운팅보다 많이 나온 단어를 negative candidate에 넣는다
- 우리의 언어모델이 실제 사람이 사용한 데이터셋의 분포와 비교하여 더 많이

Recipes for building an open-domain chatbot https://medium.com/dair-ai/recipes-for-building-an-open-domain-chatbot-488e98f658a7

Experiment: Likelihood vs. Unlikelihood

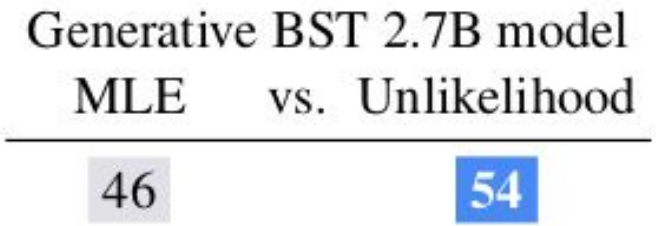


Figure 13: Self-Chat ACUTE-Eval (engagingness) MLE vs. Unlikelihood training (penalizing overexpressed n-grams). The result is not statistically significant (165 trials).

Experiment: Decoding

Minimum beam length constraint

Self-Chat Evaluations

Generative 2.7B model: Min Beam Length Constrained	vs.	Unconst.
Min. Length 5	52	48
Min. Length 10	68 **	32 **
Min. Length 20	83 **	17 **
Min. Length 40	82 **	18 **
Predictive (5,10,15,20)	69 **	31 **
Predictive (10,20,30,40)	81 **	19 **

Figure 7: Self-Chat ACUTE-Eval (engagingness) shows controlling minimum beam length gives large gains in engagingness compared to not controlling it, according to humans, with 20 being best. All rows are significant ($p < 0.01$) except the first.

Experiment: Decoding

Subsequence Blocking

Self-Chat Evaluations

Generative 2.7B model: Beam Blocking

	Block	None
3-gram Context Blocks	50	50
3-gram Response Blocks	**54**	46
3-gram Context + Response Blocks	**59**	41

Figure 8: Self-Chat ACUTE-Eval (engagingness): comparing beam-blocking variants. Blocking both context and response 3-grams during generation gives highest scores, however, none of these results are significant.

Training Data

Pre-training

- pushshift.io Reddit dataset
- Reddit의 1.5B 학습 데이터
- remove the all comments if any of the following conditions are met:
  - The author is a known bot.
  - It comes from a known non-English subreddit.
  - The comment is marked as removed / deleted.
  - It is longer than 2048 characters and does not contain spaces.
  - It is longer than 128 BPE tokens.
  - It is shorter than 5 characters.
  - It contains a URL.
  - It starts with a non-ASCII character.
  - It is further than depth 7 in the thread.

https://files.pushshift.io/reddit/

Training Data

Fine-tuning

Dataset	Trait	Paper
ConvAI2	personality, engaging	Personalizing dialogue agents: I have a dog, do you have pets too? https://arxiv.org/abs/1801.07243
Empathetic Dialogues (ED)	empathy	Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset https://arxiv.org/abs/1811.00207
Wizard of Wikipedia (WoW)	knowledge	Wizard of Wikipedia: Knowledge-Powered Conversational agents https://arxiv.org/abs/1811.01241
Blended Skill Talk(BST)	blending skills	Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills https://arxiv.org/abs/2004.08449

Experiment: BST perplexity

Model	Size	ConvAI2	WoW	ED	BST	Avg.
pushshift.io Reddit Generative	90M	18.33	31.18	14.44	18.09	20.51
BST Generative	90M	11.36	17.56	11.48	14.65	13.76
BST RetNRef	256M/90M	11.79	18.37	11.87	14.62	14.16
pushshift.io Reddit Generative	2.7B	15.70	13.73	11.06	14.36	13.71
BST Generative	2.7B	8.74	8.78	8.32	10.08	8.98
BST RetNRef	622M/2.7B	9.31	9.28	9.93	10.59	9.78
pushshift.io Reddit Generative	9.4B	15.02	12.88	10.41	13.5	12.95
BST Generative	9.4B	8.36	8.61	7.81	9.57	8.59

Table 3: **Perplexity of the pre-trained and fine-tuned models on the validation set for BST datasets.** Note that perplexity is not directly comparable between the 90M models and the larger models as 90M models use a different dictionary. Fine-tuning gives gains for each skill (task) compared to pre-training on pushshift.io Reddit alone.

Experiment: BST ACUTE-Eval

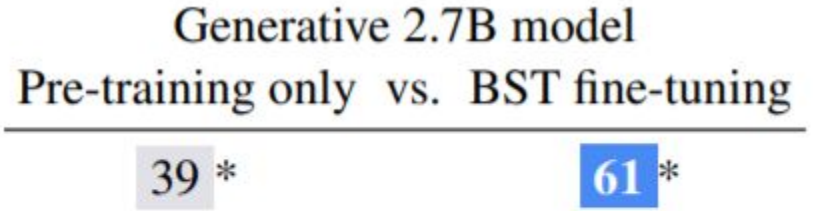


Figure 11: Self-Chat ACUTE-Eval (engagingness) shows a significant gain ($p < 0.05$) for fine-tuning on the BST Tasks.

Experiment: Comparison to Meena

	Ours	vs.	Meena
BST Generative (2.7B) std. beam	50		50
pushshift.io Reddit Generative (2.7B)	53		47
BST RetNRef (256M/90M)	60 *		40 *
BST Generative* (90M)	61 *		39 *
Wiz Generative (2.7B)	61 **		39 **
BST Unlikelihood (2.7B)	64 **		36 **
BST Generative (9.4B)	67 **		33 **
BST RetNRef (622M/2.7B)	70 **		30 **
BST Generative (2.7B)	75 **		25 **

Figure 15: Human-Chat ACUTE-Eval of **engagingness**, various models compared to Meena. Our best models are considered more engaging than Meena, rows with * ($p < 0.05$) and ** ($p < 0.01$) are statistically significant. Larger generative models with BST fine-tuning and length-controlled decoding work best.

	Ours	vs.	Meena
BST Generative (2.7B) std. beam	46		54
BST RetNRef (256M/90M)	49		51
pushshift.io Reddit Generative (2.7B)	56		44
BST Generative (90M)	59		41
Wiz Generative (2.7B)	59 *		41 *
BST RetNRef (622M/2.7B)	65 **		35 **
BST Generative (2.7B)	65 **		35 **
BST Generative (9.4B)	66 **		34 **
BST Unlikelihood (2.7B)	70 **		30 **

Figure 16: Human-Chat ACUTE-Eval of **humanness**, various models compared to Meena. Our best models are considered more humanlike than Meena, rows with * and ** are statistically significant.

Experiment: Model vs. Human-human Chat Comparisons

	Model vs. Human	
Meena (Adiwardana et al., 2020)	28 **	72 **
BST Generative (2.7B) std. beam	21 **	79 **
pushshift.io Reddit Generative (2.7B)	36 **	64 **
BST RetNRef (256M/90M)	37 **	63 **
BST Generative (90M)	42	58
BST Generative (9.4B)	45	55
BST RetNRef (622M/2.7B)	46	54
Wiz Generative (2.7B)	47	53
BST Unlikelihood (2.7B)	48	52
BST Generative (2.7B)	49	51

Figure 17: ACUTE-Eval of engagingness of models vs. humans by comparing human-bot logs to human-human logs. Rows with ** are statistically significant.

Conclusion

- BST Task Dataset
  - ConvAI, ED, WoW, BST
- Retriever(Poly-Encoder)
- Generator
- Retrieve and Refine
  - Beam search
  - unlikelihood training
  - n-gram blocking

Thank you

Reference

https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html
Personalizing dialogue agents: I have a dog, do you have pets too? https://arxiv.org/abs/1801.07243
Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset
https://arxiv.org/abs/1811.00207
Wizard of Wikipedia: Knowledge-Powered Conversational agents https://arxiv.org/abs/1811.01241
Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills
https://arxiv.org/abs/2004.08449
Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring https://arxiv.org/abs/1905.01969
The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents
https://arxiv.org/pdf/1911.03768.pdf
Towards a human-like open-domain chatbot https://arxiv.org/abs/2001.09977
https://medium.com/dair-ai/recipes-for-building-an-open-domain-chatbot-488e98f658a7
Blender Bot — Part 3: The Many Architectures
https://towardsdatascience.com/blender-bot-part-3-the-many-architectures-a6ebff0d75a6
Recipes for building an open-domain chatbot
https://medium.com/dair-ai/recipes-for-building-an-open-domain-chatbot-488e98f658a7