2023 年度 修士論文

Utilizing External and Internal Knowledge for Engaging Open-Domain Dialogue Response Generation

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

指導教員　河原 大輔 教授
研究指導名　自然言語処理研究

早稲田大学 基幹理工学研究科 情報理工・情報通信専攻

学籍番号 5121FG29

Choudhary Ritvik

July 24, 2023

Abstract

This thesis investigates the incorporation of knowledge into generative open-domain dialogue systems, a pivotal challenge in natural language processing characterized by the generation of generic and uninformed responses. Addressing the root issue of an inherent knowledge deficit, we propose two contrasting methods to enhance dialogue response quality by introducing external and internal knowledge, respectively. In the first approach, external knowledge is incorporated through a novel retriever-generator model grounded in a knowledge base of social media interactions (SMIkb). The model, comprising of a neural retriever and generator, leverages the linguistic richness of social media to augment its contextual understanding and enhance dialogue engagement and knowledge. Conversely, the second approach exploits the internal knowledge stored within the model's parameters. Utilizing the technique of knowledge distillation, a student model is trained to extract internal knowledge from two specialized teacher models or experts, representing crucial aspects of an engaging conversation: conversational ability and informativeness. This method also proves to be effective in improving the overall response engagement and informativeness. Quantitative evaluations substantiate the effectiveness of both approaches in generating improved open-domain dialogue responses. Although each approach on its own targets a knowledge source independent of the other, future research should aim to develop a comprehensive dialogue system integrating both internal and external knowledge. This study aims to be the first steps towards the advancement of natural, knowledgeable, and engaging open-domain dialogue systems.

Contents

List of Figures

List of Tables

1 Introduction

Since the advent of computing, humans have long wanted to converse with the machine in natural language. This aspiration encapsulates the desire to endow machines with the ability to understand, engage, and respond effectively in an open-ended dialogue - a feat emblematic of the complexity and spontaneity of human condition. Far from a simple execution of commands, this involves generating natural, engaging, and knowledgeable dialogue, which remains a fundamental challenge in the field of natural language processing [Adiwardana 20, Roller 21].

Nevertheless, a prevalent issue in preexisting open-domain dialogue systems is the frequent generation of highly generic ("Okay.") or uninformed ("I don't know") and uninteresting responses ("I see" etc.) [Li 16a, Xu 22]. One of the primary causes for such limitation is that, unlike humans, the models inherently lack access to rich knowledge and specific conversational intricacies during training leading to unengaging responses [Shao 17, Ghazvininejad 18].

In the light of the identified problems, recent work has proposed considering additional contextual information such as multi-turn conversation history [Zhang 18], persona [Li 16b, Cao 22], or the use of fact-based structured knowledge [Dinan 19]. Among these, our work approaches this problem from a more general standpoint of improving the overall conversational ability of generative models. Deriving inspiration from how humans learn to converse, mimicking knowledge through external social interactions and learning implicit or internal knowledge from their environment, the focus of our study is twofold. Specifically, we propose two contrasting approaches towards augmenting dialogue response quality - one utilizing an external knowledge base in a retriever-generator setup and the other distilling the internal knowledge captured within the models' parameters.

The first part is a novel approach of incorporating external knowledge through grounding a retriever-generator model in a knowledge base of **s**ocial **m**edia **i**nteractions (hereinafter referred to as SMIkb). The implementation involves employing the Dense Passage Retriever [Karpukhin 20] along with BART [Lewis 20a] seq2seq generational model trained jointly, to search the pre-indexed SMIkb, infusing relevant information with the input utterance. This approach enhances the generator models' contextual understanding, drawing from the informative and unstructured linguistic patterns prevalent in social media interactions.

Contrastingly, the second approach focuses on exploiting the internal knowledge stored in the models' parameters. Adopting the technique of knowledge distillation [Hinton 15], where a student model is trained to emulate the knowledge (behavior) held inside a more complex teacher, we first train multiple teacher models with specific abilities crucial for carrying out engaging dialogue. A target student model is then trained to generate responses by distilling the internal knowledge from two specialized teacher models, each representing a different facet of an engaging conversation. Namely, conversational ability (relevance in responses) and informativeness (or knowledgeability). Also note that this study primarily focuses on response generation for single-turn dialogues. We decided that other settings such as multi-turn cases were best addressed in future work.

In summary, we propose two novel methods to improve open-domain dialogue response generation, through incorporation of external and internal knowledge respectively. The first approach, grounded in the utilization of social media as a knowledge base, successfully boosts the quality and engagement of generated dialogue, aligning it closer to human conversational patterns. The second approach, a distillation-based method, effectively tackles the limitations of a generative dialogue model by utilizing internal knowledge from multiple teachers, offering a viable alternative in knowledge distillation. Quantitative evaluations across automatic and human metrics validate the effectiveness of both approaches.

The rest of the thesis continues with related prior work, chapter-wise explanation and evaluations of the above approaches, then finally closing with the conclusion and scope for future work in this field. We hope this study contributes towards the development of natural, knowledgeable, and engaging open-domain dialogue systems.

2 Related Work

We organize the relation of this study to existing work under the following criterion.

2.1 Dialogue Systems

Research into the development of Dialog systems have a long history starting with ELIZA [Weizenbaum 66] in the 1960s. Since then along with the advances in computation, the dialogue systems have continued to evolve at a remarkable pace. In the last decade, most open-domain dialogue systems (also referred to as “chat-bots") have come to be based on the neural network architectures originally developed for machine translation (MT) [Cho 14, Sutskever 14]. This approach of modeling conversations as a sequence to sequence (seq-2-seq) problem [Sutskever 14] became the backbone of modern generation-based dialogue systems [Vinyals 15, Sordoni 15, Serban 17]. The common theme here is the use of RNN (recurrent neural network) to model dialogue in an unsupervised setup. A seq2seq approach consists of an encoder-decoder structure where the encoder extracts the information (or context) from the input while the decoder generates the corresponding output (or response). However these simple sequence-aligned RNN-based models had a common issue of increasing memory constraints across long sequences. This, in turn, led to poor learning across distant positions [Hochreiter 01].

Hence as a result, the Transformer [Vaswani 17] utilizing self-attention to compute the contextual representations was proposed as a solution and quickly became one of the standard architecture in the field of NLP [Lakew 18, Devlin 19], and consequently also in dialogue systems [Wolf 19, Le 19, Oluwatobi 20]. Although BERT [Devlin 19], along with related developments such as RoBERTa [Liu 20] and ALBERT [Lan 20], achieved high performance on a wide array of downstream tasks, due to their nature as an encoder-only model they could not be used directly for generative tasks such as dialogue. Therefore new autoregresive encoder-decoder seq2seq models such as BART [Lewis 20a], T5 [Raffel 20] were proposed, allowing for the application for transfer learning approaches in dialogue systems. Moreover in recent months, large scale decoder-only language models (GPT-3 [Brown 20], LLaMA [Touvron 23] etc.) have also begun to be widely used for dialogue response generation.

We limit ourselves to the use of above mentioned BART [Lewis 20a] architecture as our generative model in this thesis.

2.2 Knowledge-based Generative Dialogue Models

As discussed in Section 2.1, dialogue systems have grown to become increasingly sophisticated over the years. Beyond the improvements in model architecture, incorporating additional context or external knowledge has been a field of much interest lately. Additional contexts include personal information like persona [Li 16b, Zhang 18, Cao 22], emotional information [Zhou 18, Zhang 20a], or even empathy [Rashkin 19, Tu 22]. At the same time prior works making use of external knowledge bases have become increasingly common, beginning with [Ghazvininejad 18], Wizard of Wikipedia [Dinan 19] and [Jia 20], meanwhile [Kwiatkowski 19, Chen 23] utilize their knowledge bases for a variety of question answering tasks.

The closest work to ours, in terms of including a retrieval step for generation, is [Weston 18], which proposed an approach involving pre-training the retriever and generating only over the candidates retrieved in advance from the training set. More recently [Roller 21] also tested retrieval-based dialogue generation. However, similar to [Weston 18], they utilized a retrieval model that was kept fixed during training. Our work, meanwhile, follows a different direction that does not require pre-training of the retriever but fine-tunes it along with the generator to retrieve over a much larger knowledge base of interactions at generation time.

We would also like to mention [Shuster 21], which investigates factual hallucination in dialogue retrieval-generation models with a fact-based knowledge base such as Wikipedia. Our proposed method in this thesis takes a more generalized approach, focusing solely on improving the raw conversational ability of dialogue models. Instead of factual accuracy, ours is a simple approach for generating an engaging conversation grounded in unstructured social media interactions.

2.3 Knowledge Distillation and Dialogue Response Generation

Knowledge distillation (KD) as introduced in [Hinton 15], is a key technique in machine learning, where a student model is trained to emulate the behavior (knowledge) of a (more complex) teacher model. Although initially applied to various tasks in the field of computer vision [Hinton 15, Shen 21, Feng 22], it has also been used extensively in NLP [Sanh 19, Li 21, Lee 23].

However, the use of knowledge distillation in the context of open-domain dialogue generation remains relatively rare. In recent years, [Zhang 20b] have used distillation to augment synthetic dialogue data, [Kim 21] focuses on distilling generative model responses into a retrieval model, meanwhile [Zhu 21] have proposed an approach to combining curriculum learning [Bengio 09] with knowledge distillation for dialogue generation.

This study, while in a similar space to [Zhu 21], focuses on distilling the internal knowledge from multiple specialized dialogue teachers to in-turn train a student model to generate much more engaging and informative responses.

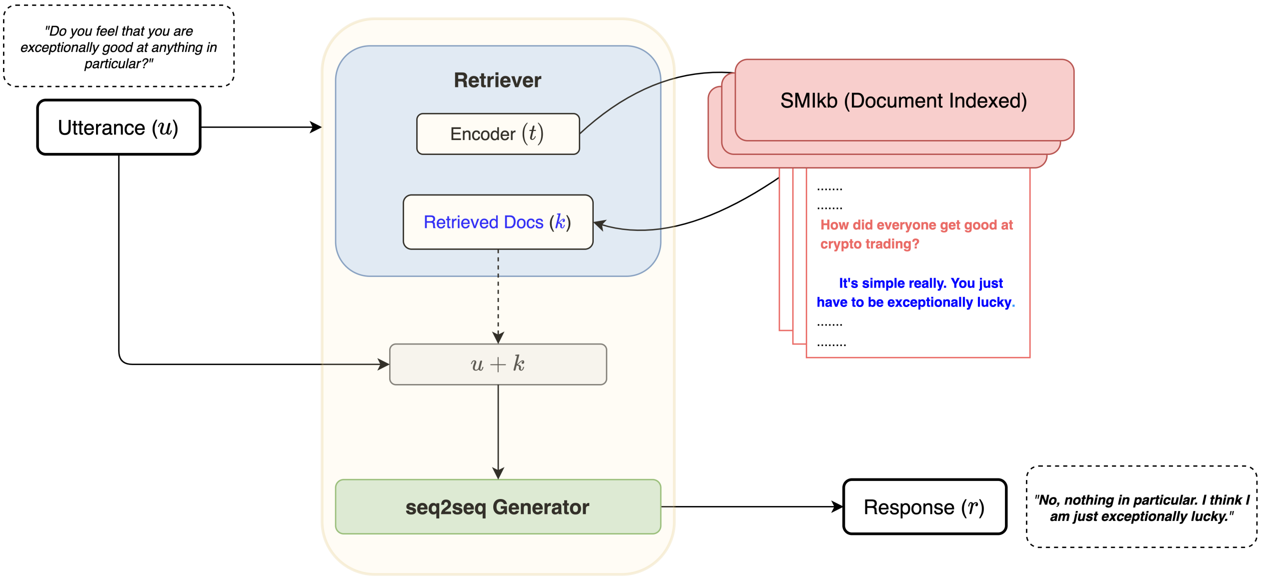

"Do you feel that you are exceptionally good at anything in particular?"
Utterance (u)
Retriever
Encoder (t)
Retrieved Docs (k)
SMIkb (Document Indexed)
....
....
How did everyone get good at crypto trading?
It's simple really. You just have to be exceptionally lucky.
....
....
$u + k$
seq2seq Generator
Response (r)
"No, nothing in particular. I think I am just exceptionally lucky."

Figure 1: Our proposed dialogue response generation approach grounded in SMIkb through a jointly trained retriever-seq2seq generator setup

3 Utilizing External Knowledge for Dialogue Response Generation

Our first proposed approach involves utilizing external knowledge, i.e. knowledge about out-of-domain topics or human conversational behavior lacking in pre-existing models, for improving the quality (relevance, engagement, and knowledge) of response generation in open-domain dialogue systems. In particular, we introduce social media interactions as an external knowledge base (SMIkb) to ground our model in for more natural and human-like response generation.

We begin with formulating the task of dialogue generation given an external knowledge source, and then proceed to explain our joint retriever-generator model as the proposed setup for utilizing the aforementioned newly developed knowledge base.

3.1 Task Formulation

Our task of response generation grounded in external knowledge can be formulated as training a model to predict a response $\mathbf{r} = (r_1, r_2, ..., r_m)$ of m words when given an input utterance $\mathbf{u}$ and a set of documents $\mathcal{D}$ that might contain relevant knowledge. We define our goal as to allow the model to learn the parameters such that when given an input utterance $\mathbf{u}$ and a knowledge base $\mathcal{D}$, the model can generate a response $\mathbf{r}$ following the probability $p(r_i|\mathbf{u}, \mathbf{r}_{<i}, \mathcal{D}; \theta)$, where θ refers to the parameters of the model.

$$\mathcal{L}(p_\theta, \mathbf{u}^{(m)}, \mathbf{r}^{(m)}) = -\log p_\theta(\mathbf{u}^{(m)} \mid \mathbf{r}^{(m)}, \mathcal{D}; \boldsymbol{\theta}) \tag{1}$$

$$= -\sum_{i=1}^{|\mathbf{r^{(m)}}|} \log p_\theta(r_i^{(m)} \mid \mathbf{u}^{(m)}, \mathbf{r}_{<i}^{(m)}, \mathcal{D}; \boldsymbol{\theta}) \tag{2}$$

3.2 Retriever-Generator Model

Inspired by recent advances in retrieval assisted QA [Guu 20, Lewis 20b], we adopt a simple joint retriever-generator setup to the task of dialogue generation. Concretely, we utilize BART, a seq2seq model pre-trained on a denoising objective, as our generative model along with the pre-trained neural Dense Passage Retriever (DPR) [Karpukhin 20] as the retriever of choice. DPR is a highly efficient neural retriever pre-trained for retrieving the top-$\boldsymbol{k}$ similar documents to an input query $\boldsymbol{u}$. It executes this by encoding both the query and the entire knowledge base through independent BERT-based encoders (as $\boldsymbol{t}$). Furthermore, we follow [Karpukhin 20] to build an offline searchable dense vector index of these embeddings for our SMIkb using the FAISS [Johnson 17] library for faster lookup. An overview of our architecture is shown in Figure 1. Application of our model to dialogue response generation can be formulated as a two-step process:

1. The retriever searching top-k documents from the pre-indexed interaction knowledge base, relevant to the input utterance.

2. The generator predicting the response to the previous utterance along with the retrieved context.

Following the notion set in Section 3.1, the probability of generating the response $\mathbf{r}$ given the utterance $\mathbf{u}$ and each of the top-k documents d_j from the knowledge base $\mathcal{D}$ can be defined as

$$p(\mathbf{r}|\mathbf{u};\theta,\lambda) = \sum_{j}^{k} p_{\lambda}(d_j|\mathbf{u};\lambda) \prod_{i} p_{\theta}(r_i|\mathbf{u},\mathbf{r}_{<i},d_j;\theta), \tag{3}$$

where θ and λ are parameters for the generator and retriever, respectively. They are both fine-tuned jointly in an end-to-end fashion, with the retriever providing additional context that is concatenated together with the input at the time of generation. As there is no “correct” document source in the knowledge base, we consider it to be a latent variable. Therefore, during decoding we marginalize these probabilities over all the retrieved documents to return the most probable (best) response using beam search.

3.3 Experiments

We evaluate our proposed retriever-generator model grounded in multiple external knowledge datasets spanning various domains, on the task of open-domain dialogue generation. The results are then compared against two competitive BART-based baselines.

3.3.1 SMIkb: Knowledge base of Unstructured Social Media Interactions

Aiming to improve the raw communication ability of dialogue systems by mimicking human response behavior, we developed our own human-readable external knowledge base, built primarily of unstructured human-human social media interactions (SMIkb).

It comprises of entries from top thread titles and their top 100 comments from Reddit, an American social news aggregation and discussion site, throughout 2020 (January-November). For our

SMIkb	
title	text
LPT: If you borrow something like a tool or a generator from someone, return it in BETTER shape than you got it.	My dad always said that returning something in the same condition you received it is the absolute bare minimum.
SoftBank Nears $40 Billion Deal to Sell Arm Holdings to Nvidia	Nvidia is priced decentlyfor what they offer.
Apple to Give Employees Paid Time Off to Vote in U.S. Election	This exactly. A large majority of disenfranchised communities work jobs that don't observe federal holidays.
Apple may be working on a foldable iPhone	I can confirm that Apple would be stupid to not be working on one. Whether they ever release one is up for debate, but they're definitely working on one.
Anyone else feel like there are so many good games that are completely spoiled by a single bad mechanic?	I don't have an issue with all inventory management, I have an issue with limited capacities. I'd much rather be able to pick it all up and sort it later in town. At least mods can fix that.
Google changed my device trade in value from $350 to $17.50.	Isn't it crazy we need good Samaritans to step in and help with these things because Google CS can't?

Table 1: Snapshot of SMIkb

study, we used the Pushshift API [Baumgartner 20], an active big-data project maintaining copies of entire reddit data, to scrape over 324 different sub-reddits emulating a wide information source totaling to 1,639,543 unique entries. From this collection, a random selection of 600,000 (due to memory limitations) makes up our SMIkb. A snapshot of the same is shared in Table 1.

Furthermore, to verify the effectiveness of using a conversational knowledge base like Reddit, we compared ours to a pure Wikipedia knowledge base (ref. "Wiki") of the same size (random sample of 600k entries) containing the wiki page title and the leading 100 words. Additionally, we also tested a 1:1 combination of the above two bases (ref. "Mix").

3.3.2 Fine-tuning datasets

For end-to-end fine-tuning on our open-domain dialogue generation task, we use a combination of datasets of varying nature from different sources.

Open-domain dialogue datasets (ODD) For end-to-end fine-tuning on the dialogue generation task, we use a combination of datasets of varying nature from different sources. The first is DailyDialog [Li 17], a high-quality, human written and annotated dataset covering various topics from daily life over a span of 13,118 conversations. Next, we add DailyDialog++ [Sai 20], a dataset comprising of multiple relevant responses for over 19,000 different contexts. In addition we also include the Cornell Movie-Dialogs Corpus [Danescu-Niculescu-Mizil 11], which is a corpus of transcription of movie dialogues. Although originally multi-turn, for the purpose of this study, all of these were extracted and converted into a series of single-turn conversations. We split our set into *train*, *valid*, and *test* with a ratio of 70%, 15%, and 15% respectively. The overall breakdown is given in Table 2.

Comments from Reddit In addition to the publicly available natural open domain data sets discussed in the previous section, we further extract another 200,000 comment pairs from Reddit, distinct from SMIkb. They act as pseudo dialogue pairs to supplement our knowledge base. Preliminary experiments including them in the mix showed minor improvements, hence this step might

Dataset	Total (turns)	Train	Valid	Test
DailyDialog	76,743	53,721	11,511	11,511
DailyDialog++	39,913	27,939	5,987	5,987
Cornell Movie-Dialogs	221,088	154,762	33,163	33,163
Reddit (pseudo extracted)	200,000	140,000	30,000	30,000

Table 2: Overview of datasets in use for retriever-generator model setup

help training by bridging the distribution gap between the external knowledge base and the natural fine tuning dialogue data.

3.3.3 Experimental Setup

Implementation Details Our joint retriever-generator model consists of a pre-trained Dense Passage Retriever and BART-large (24 layers, 406M), which are later fine-tuned together on SMIkb and dialogue datasets. The model is trained mostly with the default parameters, batch size of 1, and an initial learning rate of 3×10^{-5}. We further experiment with various values of k for our top-k document retrieval, while beam search with size of 5 is used as our response decoding strategy. Fine-tuning is performed with an Nvidia V100 GPU using the HuggingFace library [Wolf 20] along with Pytorch Lightning [Falcon 19].

Baselines We consider two strong baselines based on a vanilla BART-large with no retriever to investigate the effectiveness of our approach.

1. The first is fine-tuned solely on the datasets mentioned in Section 3.3.2 (ref. “Baseline 1”) with no SMIkb.
2. Next to confirm the effectiveness of our providing external data through our retriever-generator setup, we merge the entire SMIkb interactions into our training data, and simply fine-tune the vanilla model on this new extended set. (ref. “Baseline 2”).

Note that although we choose BART as our generator and baseline for its size and relative ease in training, our proposed SMIkb based modeling setup could possibly also be extended to larger models.

3.4 Evaluation

To measure the impact of social media interactions, the generated responses were evaluated through both automatic and human evaluations. The results are compiled in Tables 3 and 4.

Model Setup	Training Data	Knowledge Base (Retrieval)				BLEU-4	Dist-1	Dist-2			
Baseline 1	ODD	None				1.31	0.20	0.96			
Baseline 2	ODD + SMIkb	None				1.05	0.12	0.47			
			$k = 3$			$k = 5$			$k = 7$		
			BLEU-4	Dist-1	Dist-2	BLEU-4	Dist-1	Dist-2	BLEU-4	Dist-1	Dist-2
Ours (SMIkb)	ODD	SMIkb	**9.78**	**2.80**	**16.90**	**10.51**	**5.50**	**26.63**	**10.48**	**5.51**	**26.62**
Ours (Wiki)	ODD	Wiki	6.93	2.57	14.91	7.14	4.94	23.38	7.11	5.02	23.79
Ours (Mix)	ODD	SMIkb + Wiki	6.03	2.45	14.08	6.20	4.71	22.25	6.21	4.71	22.23

Table 3: Automatic evaluation of generated responses across various values of k for top-k retrieval

Model Setup	Human Eval.		
	Relevance	Engagement	Knowledge
Gold (Test-Data)	3.50	3.33	3.47
Baseline 1	2.82	2.35	3.00
Baseline 2	3.03	3.02	2.89
Ours (SMIkb)	**3.84**	3.75	3.60
Ours (Wiki)	3.40	3.75	**3.76**
Ours (Mix)	3.62	**3.80**	3.71

Table 4: Human evaluation of generated responses for the best $k = 5$

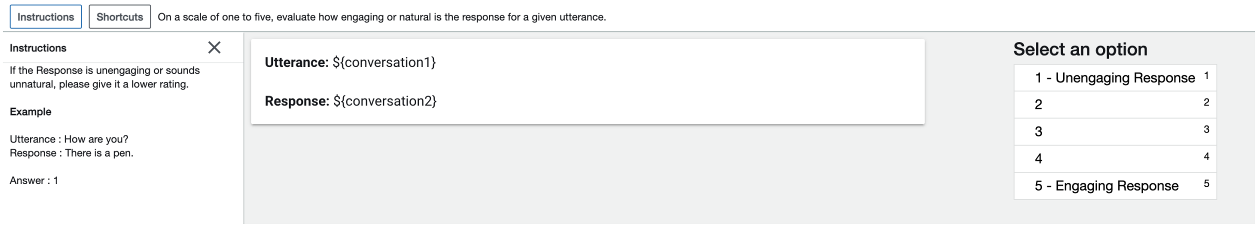
3.4.1 Automatic Evaluation

We perform a series of automatic evaluations on the jointly fine-tuned seq2seq model. First, to measure the relevance between the generated output and gold response we calculate the extent of their n-gram overlap through the BLEU [Papineni 02] score metric. Next, to measure the diversity of the responses generated the Distinct metric [Li 16a] is used. Distinct-N is a measure of unique n-grams as a fraction of total generated words. We calculate both the unigram and bigram values for our models.

3.4.2 Human Evaluation

We would also like to note that it has been widely reported that automatic evaluation metric do not correlate well with the actual quality of the generated outputs [Liu 16]. Thus, we additionally performed human evaluation of the responses with the highest BLEU ($k = 5$) through Amazon Mechanical Turk on a 5-point Likert scale to further measure the effectiveness of our model. The evaluation is carried out regarding three different markers connecting back to the primary goal of our work:

- **Relevance**: How well does the generated response correlate with the input utterance?
- **Engagement**: Does the generated response makes the conversation engaging or how likely are the evaluators to continue the conversation?
- **Knowledge**: How knowledgeable or sensible is the generated response?



On a scale of one to five, evaluate how engaging or natural is the response for a given utterance.

Instructions

If the Response is unengaging or sounds unnatural, please give it a lower rating.

Example

Utterance : How are you?
Response : There is a pen.

Answer : 1

Utterance: ${conversation1}

Response: ${conversation2}

Figure 2: Screen shown to raters on MTurk for evaluating the response on a 5-point Likert scale

For carrying out the human-evaluation of the generated responses, we selected our evaluators to be English speakers from the United States with an approval rate of over 90%. The average pay was set at 1 cent per question. The evaluators were asked to score 100 responses selected at random from the test set, on a scale of 1-5. Each response was scored by 7 different evaluators, and their average was calculated. A snapshot of the Mturk evaluation screen shown to the raters is shown in Fig. 2

3.5 Results and Analysis

First, with automatic evaluation, we observe that our method of introducing social interactions through a retriever at generation time maintains task performance and allows for a more diverse set of responses, as shown with an increase in all of the measured metrics over both the baselines. The baselines do not have a retrieval step and therefore do not have an effect due to changing k. **bold** refers to the best scores across all k among the generated responses, while ODD is the collection of **O**pen-**D**omain **D**atasets from Section 3.3.2.

Next, our Reddit-based SMIkb model outperformed other combinations, confirming an increase in response quality. Furthermore, experiments comparing the effect of top-k on generation quality showed $k = 5$ as the sweet spot with the highest BLEU, which was then evaluated further in human evaluation.

Similar to the automatic evaluation, all of the proposed combinations showed improvements over the baseline in human evaluation. In line with our proposed approach, the SMIkb model recorded the highest score in terms of conversational relevance. Meanwhile, the mix of SMIkb and Wikipedia seems to find the appropriate balance between conversational ability and "knowledgeability" with the metrics even exceeding the gold score. We hypothesize that this is not just due to high generation quality but also the nature of our task, which might make our model responses more relevant to a worker when assessed in a single-turn context. Overall, the results verified our proposed approach of learning directly from social media interactions leading to large improvements over the baseline in all metrics of interest. Furthermore, significance tests of bootstrap re-sampling showed that the above evaluation score differences between baselines and our models were statistically significant ($p < 0.01$). Examples of the various generated responses across all the models are shared in Table 5.

In addition, we looked at cases where our model struggles or fails at generating a natural response, a select few of which are compiled in Table 6. In some of these particular cases, the baseline generations, although generic, turn out to be more relevant. We believe some of these long

Input Utterance	SMIkb	Wiki	Mix	Baseline
Here comes the birthday cake.	Wow, what a lovely cake. I'll have a piece.	Wow, what a beautiful cake. It must have tasted as good as I expected.	Really? Wow. Happy birthday to you! You look so healthy.	Thank you.
I'd better go. I gotta get up early. Go over to my parents'.	What are you going to do for breakfast?	You mean you don't wanna go with me?	Why don't you stay at my place? I'll cook dinner.	I don't think so.
Something's burning?	Oh my goodness, it's on fire!	Oh my goodness, it ' s burning!	Oh, my bread! It's burning!	No, it's not.
Would you mind if we just called it an early night?	Even I couldn't. I have work tomorrow.	I'd love to. But I've got to work tomorrow.	What are you talking about? We just ate?	I don't think so.
Isn't it wonderful walking here?	What do you mean? Was it bad?	Not at all. The air is fresh, the walk is easy and the sunshine is kind.	Yes, you can see the whole city from here.	Yes, it's very beautiful.

Table 5: A comparison of generated responses, grounded in various external knowledge bases

Input Utterance	SMIkb	Wiki	Mix	Baseline
Who was the conductor?	Oh speaking of the conductor, he was indeed a genius. A descendent of Chinese.	I don't remember the name.	Oh, speaking of the conductor, he did an excellent job.	I know.
Oh god you scared me!	Ahahah! We need a fitness instructor.	I'm sorry.	Sorry. I was trying to scare you.	Is something wrong?

Table 6: Selected examples where the proposed retriever-generator setup struggles or fails at generating a relevant response

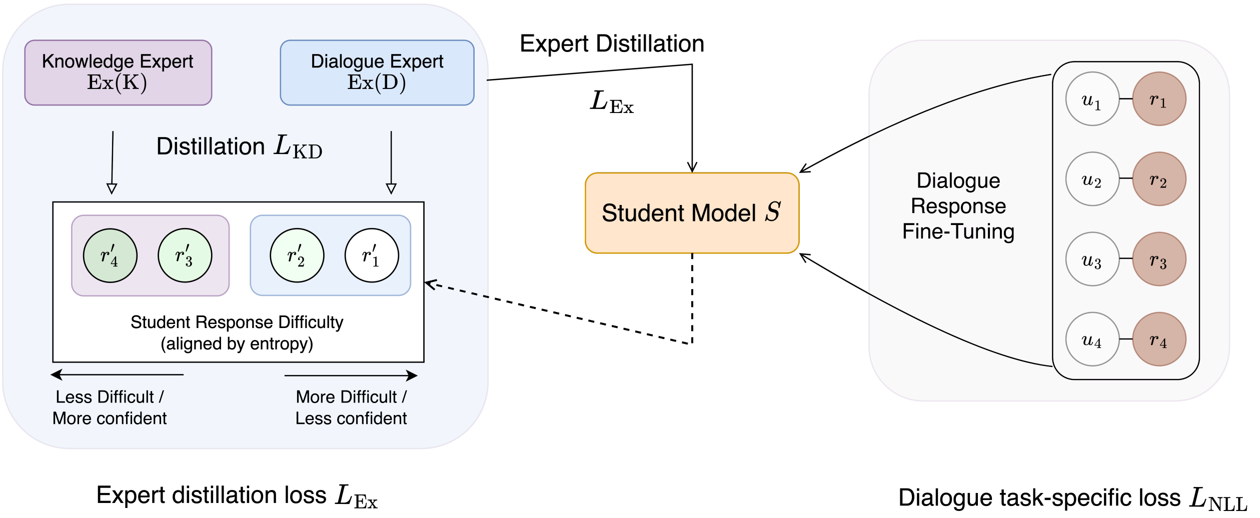


Figure 3: Overview of the proposed multi-expert distillation framework for response generation

responses with unrelated information to be an artifact of our generation model being overly dependent on the knowledge base. While social media may simulate human-like conversations in a large variety of situations, it is still far from being a perfect stand-in for real-life dialogue. Therefore, our future work in this direction should look at not only the quality and scope of the knowledge base, but also consider selecting *when* to ground and make use of the said knowledge during response generation.

4 Distilling Internal Knowledge for Dialogue Response Generation

We further address the previously discussed problem of pre-existing models' failure to capture knowledge and subsequent drop in engaging responses, by presenting an alternate, simple yet effective approach. In contrast to utilizing an external knowledge source for context, we turn our attention inwards, towards internal knowledge held inside a model's parameters.

We first redefine an engaging conversation to be made up of various abilities, training multiple expert models for each. The proposed method then aims to train a student model by distilling the relevant knowledge from expert models, in-turn allowing our target student to effectively learn the specialized abilities required for a more engaging dialogue response generation.

4.1 Task Formulation

Unlike Section 3.1, this alternate approach does not inherently require an external context source, therefore the loss to minimize is the standard negative log-likelihood for dialogue response generation. Hence, given the input utterance $\mathbf{u}$, the model can generate a response $\mathbf{r}$ following probability $p(r_i|\mathbf{r}_{<i}, \mathbf{u}; \theta)$.

$$\mathcal{L}(p_\theta, \mathbf{u}^{(m)}, \mathbf{r}^{(m)}) = -\log p_\theta(\mathbf{u}^{(m)} \mid \mathbf{r}^{(m)}; \boldsymbol{\theta}) \tag{4}$$

$$= -\sum_{i=1}^{|\mathbf{r^{(m)}}|} \log p_\theta(r_i^{(m)} \mid \mathbf{u}^{(m)}, \mathbf{r}_{<i}^{(m)}; \boldsymbol{\theta}) \tag{5}$$

4.2 Response Generation with Multi-Expert Distillation

As briefly mentioned earlier, in particular, we focus on two significant abilities that contribute to an engaging response: (1) conversational ability (relevance) and (2) informativeness (knowledgeability). Correspondingly, we first train two models, namely **Ex(D)** and **Ex(K)**, on the respective conversational and informative domain dialogue data, following the loss described in Section 4.1. These models act as our *experts* for teaching our final student through knowledge distillation (KD).

Note that an engaging conversation may consist of various aspects and abilities not limited to conversational ability and knowledge, but also include coherence, personality, topicality and others. However, in the scope of this study, we intentionally limit ourselves to conversational ability and informativeness as we consider these two to be the fundamental abilities that initially make up an engaging dialogue system. Furthermore, these abilities can also be considered, to an extent, a superset that has some overlap (a high conversational ability overall can be linked to high relevance and coherence, while high informativeness can be a starting point for topical interestingness or elements of persona) with the other finer aspects of an engaging conversation. Nonetheless, given enough time and compute, our proposed framework can also be extended to include learning from multiple detailed teachers specializing in each of the above mentioned aspects.

As introduced in [Hinton 15], given a specified teacher and a student model, KD aligns the student to the teacher by minimizing the Kullback-Leibler (KL) divergence between their output probability distributions. The idea is the student model will learn to predict the teachers ‘soft’ distribution - output probability across all classes, instead of only the final predicted class, thereby generalizing the knowledge *internal* to the teacher. For student S and teacher T predictions, say $\mathbf{r_s}$ and $\mathbf{r_t}$, KD minimizes the following distance.

$$\mathcal{L}(\mathrm{T})_{\mathrm{KD}} = \mathrm{KL}(\mathbf{r_s} \mid\mid \mathbf{r_t}) \tag{6}$$

Now, when in the presence of multiple experts, it becomes crucial for the student to optimize the knowledge learned from each expert. We address the same by taking inspiration from prior work in the field of curriculum and active learning [Bengio 09, Settles 08] where we consider student difficulty as a barometer for effectively adjusting the experts' distillation signal during training. [Bengio 09] formulates curriculum learning as a method to gradually increase the complexity of examples which has shown to be helpful for training a more generalized model. Therefore, based on related previous work [Li 21, Bengio 09], we adopt student entropy H over its response as our preferred metric for assessing a sample's complexity/difficulty (or confidence; low confidence implies high difficulty).

$$H = -\sum_i p(r_i|\mathbf{r}_{<i}, \mathbf{u}) \log p(r_i|\mathbf{r}_{<i}, \mathbf{u}) \tag{7}$$

Using the above student response difficulty metric, we realign the examples for training from easiest to hardest, and the distillation signal from each expert is then divided across the batch accordingly. For difficult examples, the student $\mathbf{r_s}$ may benefit from an easy-to-learn expert **Ex(D)**, while the other expert may help with the easier examples instead. We formulate such a division of expert distillation signals as follows.

$$\mathcal{L}_{\text{Ex}} = \gamma\ \mathcal{L}(\text{Ex(D)})_{\text{KD}} + (1-\gamma)\ \mathcal{L}(\text{Ex(K)})_{\text{KD}}, \tag{8}$$

where γ is a parameter for balancing the experts. Combining the two processes with an adjustable parameter λ, we can define the overall loss $\mathcal{L}$ to optimize our dialogue response generation model.

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda\ \mathcal{L}_{\text{Ex}} \tag{9}$$

An overview of our proposed multi-expert distillation framework is shown in Fig. 3. The target student model S is trained on a combination of distillation loss from multiple experts of specific abilities $\mathcal{L}_{\text{Ex}}$ and a dialogue response generation task-specific loss $\mathcal{L}_{\text{NLL}}$.

4.3 Experiments

We train, distill, and evaluate our multi-expert distillation framework on a mixture of existing open-domain dialogue datasets. The results are then compared against various seq2seq BART baselines.

4.3.1 Experimental Setup

Our expert and student models discussed in the previous sections consists of BART [Lewis 20a] seq2seq models fine-tuned on various dialogue and knowledge datasets. The implementation details are similar to that of Section 3.3.3, with the number of GPUs increased to 4.

4.3.2 Datasets

We fine-tune our expert and student models on various open-domain dialogue datasets.

Open-domain dialogue datasets (ODD) We fine-tune our student on a similar combination of ODD as mentioned in Section 3.3.2. However, the total number of training samples for each model is set at 200,000.

Model	Arch (BART)	Training Data	Turns
Ex(D)	Large	ODD	200,000
Ex(K)	Base	WikiDialog	500,000
Ex(K_S)	Base	Reddit	500,000

Table 7: Training data for specialized teachers

Knowledge-based (pseudo) dialogue datasets For training our specialized expert models we use the following data.

For our informativeness (knowledge) expert **Ex(K)**, we use WikiDialog [Dai 22] as our dataset of choice. WikiDialog is a synthetically generated dialogue dataset extracted from converting Wikipedia passages into a conversation between two pseudo speakers. Furthermore, we also consider the Reddit-based social media dialogue dataset (SMIkb) introduced in Section 3 as an alternative informative dataset to train a different expert model **Ex(K_S)** for comparison. We set the total number of training samples for informative models to 500,000, as this is considered a more difficult task to train. A summary of the expert training data is compiled in Table 7.

4.4 Specialized Teachers

As discussed in Section 4.2, we train two expert models that act as specialized teachers for our student model:

1. **Ex(D)** for conversational ability (relevance)
2. **Ex(K)** for informativenes (knowledgeability)

In particular, **Ex(D)** is a BART-large model trained on a set of open-domain dialogue datasets, while **Ex(K)** is a BART-base[1] model trained on more informative Wikipedia-extracted pseudo conversational pairs.

Note that the latter **Ex(K)** is trained on a more information dense dataset than the simple daily-life dialogues of the former **Ex(D)** model. Furthermore, to quantify this difficulty of the informative datasets compared to the other, we randomly select 100 utterance-response pairs and calculate their average perplexity (PPL) as generated by a language model similar in size to our target, GPT-2 (117M). The results are compiled in Table 8.

Following the above observed trend, we can consider the Wikipedia or Reddit-based knowledge/information dataset to be sufficiently diverse and contain uncommon or noisy conversational structures compared to casual daily-life interactions, and hence we can consider our expert model trained solely on the same, the informativeness expert **Ex(K)** to be a more difficult to learn teacher for a student as compared to the other daily-life dialogue teacher **Ex(D)** which could be relatively easy to learn for our student model due to an easier vocabulary and the overlap with the NLL training data of the student itself. To back this up, we also share a few example pairs of training samples from each of the datasets mentioned in Section 4.3.2 in Table 9.

[1] BART-base was selected for its relative ease and stability of training.

Dataset	Avg. Perplexity (↓)	
	Utterance	Response
ODD	195.56	60.90
WikiDialog	240.49	306.42
Reddit	179.27	309.10

Table 8: Average perplexities of utterances and responses across various datasets

Dataset	Example pairs from the set
ODD	*Utt*: Hey, Jenny. Would you like to go to dinner with me? *Resp*: I don't know. You know what they say about office romances.
	Utt: Waiter, can I have the bill please? *Resp*: Wait a moment. It's $30.
WikiDialog	*Utt*: What is Selo Zyuzino? *Resp*: Zyuzino, Moscow Oblast Zyuzino is a rural locality (a "selo") in the Ramensky District in Moscow Oblast, Russia.
	Utt: What is known about the Karuna Trust? *Resp*: Karuna Trust (Sri Lanka) is a voluntary non-profit organization dedicated to improving the living standards of materially poor people in Sri Lanka.
Reddit	*Utt*: SoftBank Nears $40 Billion Deal to Sell Arm Holdings to Nvidia *Resp*: Nvidia is priced decently for what they offer.
	Utt: Apple may be working on a foldable iPhone. *Resp*: I can confirm that Apple would be stupid to not be working on one.

Table 9: Example conversation pairs from varying dialogue datasets

Building upon the above reasoning, recall that when learning the distillation signal, we choose the expert based on the student's measure of difficulty of a training sample. Hence over the course of training, the samples deemed difficult by the student are matched with an easier-to-learn teacher, and the opposite also follows. For our experiments, the multiple experts are balanced equally with $\gamma = 0.5$.

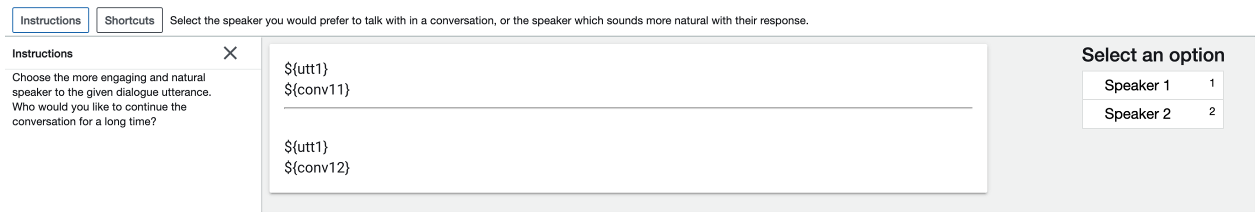
4.5 Evaluation

The responses generated on the ODD test set were evaluated against a combination of baselines to measure the effectiveness of our proposed setup. Due to differences in training conditions among expert models, we limit comparisons to similar-sized baselines.

As in Section 3.4, the models were evaluated through both automatic and human evaluations.

4.5.1 Baselines

For our evaluations we train and compare the proposed model against five different baselines as follows.



Select the speaker you would prefer to talk with in a conversation, or the speaker which sounds more natural with their response.

Choose the more engaging and natural speaker to the given dialogue utterance. Who would you like to continue the conversation for a long time?

${utt1}
${conv11}

${utt1}
${conv12}

Figure 4: Screen shown to evaluators on MTurk for performing pair-wise comparison of responses

1. **Vanilla**: First is a vanilla BART-base model trained directly on the 200k samples from the ODD mentioned in Section 4.3.2. This is in direct comparison to our similar-sized target student model.

2. **Vanilla -** ***ALL***: Next, to confirm the effectiveness of distillation over simple fine-tuning on a combination of two datasets, we train a vanilla BART-base on a total of 700k samples by merging all the datasets.

3. **Ex(D)-Distill** and **Ex(K)-Distill**: To verify the advantage of our multi-expert setup over regular distillation, we compare two baselines trained with either of our experts (equivalent to standard knowledge distillation).

4. **Multi Expert w/ Ex($\mathbf{K_S}$)**: To measure the effect of using a Wikipedia-based model as our knowledge expert, we train a SMIkb-based (cf. Section 3.3.1) expert model Ex(K_S) and compare this related variation (ref. ME-Ex (K_S)) of our proposed setup.

Also note that while we select BART due to size limitations, our setup could easily be extended to larger models.

4.5.2 Automatic Evaluation

Following previous literature in this domain, we measure BLEU [Papineni 02] and perplexity (PPL) [Melis 17] scores as a stand-in for the overall quality of our responses.

4.5.3 Human Evaluation

As discussed in Section 3.4.2, automatic evaluations metrics alone do not sufficiently evaluate the actual engagement of generated outputs [Liu 16]. Therefore, following contemporary work [Roller 21], for this approach we perform human evaluation by pairwise comparison of the generated responses of our proposed method and respective baselines. The evaluation is carried out via Amazon Mechanical Turk on two primary metrics:

1. **Engagement**: Whether the human evaluator would prefer to continue the conversation with this agent for a long period of time.

2. **Informativeness**: Whether the response seems sufficiently informative or knowledgeable.

Model	Distill	Training	Automatic Eval.		Human Eval. (***Ours*** vs. Model, %)			
(BART)	Expert	Data	BLEU-4 (↑)	PPL (↓)	Engagement		Informativeness	
					Win	Lose	Win	Lose
Vanilla	-	ODD (200k)	1.56	190.4	**83**	17	**67**	33
Vanilla - *ALL**	-	ODD (200k) + Wiki (500k)	**2.93**	182.1	**57**	43	**51**	49
Ex(D) Distill	Ex(D)	ODD (200k)	1.64	33.55	**79**	21	**78**	22
Ex(K) Distill	Ex(K)	ODD (200k)	1.85	44.67	**75**	25	**84**	16
Ours ME-Ex (K_S)	Ex(K_S) + Ex(D)	ODD (200k)	1.71	**30.60**	**79**	21	**84**	16
Ours	Ex(K) + Ex(D)	ODD (200k)	1.88	35.37	-	-	-	-

Table 10: Automatic and human evaluation of the generated responses by multi-expert distillation framework against baselines

The evaluators were asked to compare 100 responses from the test set, seven evaluators saw each response, and their win/loss percentages were calculated. We selected our evaluators to be English speakers from the United States with an approval rate of over 90%. A snapshot of the Mturk screen shown to the evaluators is shown in Fig. 4

4.6 Results

The evaluation results are summarized in Table 10 with the best scores in **bold**.

First, with automatic evaluation, we observe our proposed model generally outperforms most baselines with a slight drop in BLEU and increased perplexity. We hypothesize this as an artifact of our method due to the increased diversity in the engaging responses. However, as discussed earlier n-gram-based metrics are not an entirely accurate representation of response quality.

Human evaluation of our proposed multi-expert knowledge distillation framework, meanwhile, showed notable improvements against all of the above baselines. In line with our initial hypothesis, the proposed approach won over the baselines in terms of generating responses that are shown to have higher engagement and informativeness. Also note that although evaluated on the same test set, Vanilla-*ALL** is a baseline directly fine-tuned on a much larger set of training data (700k total vs 200k for the proposed method). Hence while not entirely an equal comparison, Vanilla-*ALL** acts as a highly competitive baseline while still falling slightly short of the proposed method.

Moreover, the human evaluators preferred our multi-expert approach to both the single-expert baselines in pairwise comparison, hinting that both abilities, conversational ability and informativeness or knowledgeability, play a crucial role in generating more engaging responses in the scope of open-domain dialogue. However it must be noted that due to the nature of our proposed method, utilizing the parametric knowledge stored inside in the models, a majority of the learning process as well as other factors that impact response quality are effectively under the blackbox of language models. Further insights into specific elements that make up and affect the engagement of an open-domain system as well as overall explainability of the student model currently remains as an unsolved problem.

A set of the responses generated by the proposed method and others is compiled in Table 11.

Input Utterance	Proposed Method	ME-Ex(K_R)	Vanilla Baseline
Removal Point? What're you talking about!	It's Removal Point. That's where they put the boat.	I'm not talking about anything. I'm talking	I'm not talking I'm not talking
I know why you broke up with me in college. Can you forgive me?	I know, but I cannot forgive you.	I don't forgive you.	I don't know what you are talking about.
I live in Haitian District where there are lot of universities .	That's great! What kind of universities do you like in Haiti?	What kind of universities are you looking for?	Do you have a good education?
Shes very striking, don't you think? Are these glasses a fashion where she comes from?	No, I don't think so. Its probably just fashion.	No. I don't think so. She's a very beautiful girl.	I don't know. I don't think so.
That sounds fine. Why did you choose to apply to our company?	I wanted to work in a company where I could learn a lot.	I decided to apply to our company because I want to get a better job.	Because I am interested in your company, I decided to apply to your company.

Table 11: A snapshot comparison of responses generated from multi-expert distillation models and the baseline

5 Conclusion

Through the study we aim to solve the fundamental problems of low engagement and informativeness in the responses by preexisting mid-sized generative dialogue models.

We attempt to solve the root cause of inherent lack of knowledge by proposing a twofold contrastive approach for incorporating knowledge into a dialogue system. The first aims to improve the conversational ability of dialogue systems by grounding the responses in an *external* knowledge base made of human-like social media interactions (SMIkb). We utilize the neural retriever-generator model fine-tuned jointly, capable of retrieving relevant knowledge at the time of generation to assist a more engaging and knowledgeable dialogue response. Automatic and human evaluations against various baselines confirmed the effectiveness of our approach. The second approach, in contrast, looks towards harnessing the knowledge held *internally* in a model's parameters for a more engaging dialogue response generation. By distilling knowledge from multiple experts, we supplement our target student model abilities to generate a more engaging and informative response. The improvements were later confirmed with human evaluators.

Although overall we have looked at knowledge incorporation for dialogue generation in two independent settings, internal or external, it should come as no surprise that humans, on the other hand, utilize both sources seamlessly at all times. Therefore, the development of a comprehensive knowledgeable dialogue system, intelligently utilizing different forms of knowledge during generation while maintaining factuality and coherence, remains as important future work in this direction.

Acknowledgement

約三年間にわたる河原研究室での研究生活において、研究方針の決定、実験設定の構築、そして論文の執筆含めて、研究そのものとの向き合い方を教示して下さった河原先生に心より感謝の意を表します。変化の多い期間でありましたが、それにもかかわらず、自然言語処理という一つの分野にここまで興味を抱くことができたのも、ひとえに先生のご指導のおかげです。重ねて、感謝申し上げます。まだ NLP の道に突き進んでいきたいと思いました。

次に、この 3 年間で、ホワイトボードの前での議論から軽井沢やアメリカまで、支えとなってくれた一期生の同期と優秀な後輩たちに感謝いたします。本当にありがとうございました。

I would like to thank my parents and my sister, for their never-ending support over the past six years. This thesis has only been possible thanks to their countless sacrifices and continued love and support throughout my entire university journey in Japan.

I would also dedicate a part of this thesis to my late best friend Isha, without whom grad school or this part of my life would never have happened. Thank you and rest in peace.

Finally, thank you to all my friends and everyone who supported me this far.

References

[Adiwardana 20] Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al., “Towards a human-like open-domain chatbot”, arXiv preprint arXiv:2001.09977 (2020)

[Baumgartner 20] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J., “The Pushshift Reddit Dataset”, in Choudhury, M. D., Chunara, R., Culotta, A., and Welles, B. F. eds., Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020 , pp. 830–839, AAAI Press (2020)

[Bengio 09] Bengio, Y., Louradour, J., Collobert, R., and Weston, J., “Curriculum Learning”, in Proceedings of the 26th Annual International Conference on Machine Learning , ICML ’09, p. 41–48, New York, NY, USA (2009), Association for Computing Machinery

[Brown 20] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D., “Language Models are Few-Shot Learners”, in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. eds., Advances in Neural Information Processing Systems , Vol. 33, pp. 1877–1901, Curran Associates, Inc. (2020)

[Cao 22] Cao, Y., Bi, W., Fang, M., Shi, S., and Tao, D., “A Model-agnostic Data Manipulation Method for Persona-based Dialogue Generation”, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 7984–8002, Dublin, Ireland (2022), Association for Computational Linguistics

[Chen 23] Chen, W., Verga, P., Jong, de M., Wieting, J., and Cohen, W. W., “Augmenting Pre-trained Language Models with QA-Memory for Open-Domain Question Answering”, in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics , pp. 1597–1610, Dubrovnik, Croatia (2023), Association for Computational Linguistics

[Cho 14] Cho, K., Merriënboer, van B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y., “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pp. 1724–1734, Doha, Qatar (2014), Association for Computational Linguistics

[Dai 22] Dai, Z., Chaganty, A. T., Zhao, V. Y., Amini, A., Rashid, Q. M., Green, M., and Guu, K., “Dialog inpainting: Turning documents into dialogs”, in International Conference on Machine Learning , pp. 4558–4586PMLR (2022)

[Danescu-Niculescu-Mizil 11] Danescu-Niculescu-Mizil, C. and Lee, L., “Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.”, in Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011 (2011)

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) , pp. 4171–4186, Minneapolis, Minnesota (2019), Association for Computational Linguistics

[Dinan 19] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J., “Wizard of Wikipedia: Knowledge-Powered Conversational Agents”, in International Conference on Learning Representations (2019)

[Falcon 19] Falcon, W. and The PyTorch Lightning team, , “PyTorch Lightning” (2019)

[Feng 22] Feng, P. and Tang, Z., “A Survey of Visual Neural Networks: Current Trends, Challenges and Opportunities”, Multimedia Syst. , Vol. 29, No. 2, p. 693–724 (2022)

[Ghazvininejad 18] Ghazvininejad, M., Brockett, C., Chang, M.-W., Dolan, B., Gao, J., Yih, W.-t., and Galley, M., “A knowledge-grounded neural conversation model”, in Proceedings of the AAAI Conference on Artificial Intelligence , Vol. 32 (2018)

[Guu 20] Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W., "REALM: Retrieval-Augmented Language Model Pre-Training", in Proceedings of the 37th International Conference on Machine Learning , ICML'20, JMLR.org (2020)

[Hinton 15] Hinton, G., Vinyals, O., and Dean, J., "Distilling the knowledge in a neural network", arXiv preprint arXiv:1503.02531 (2015)

[Hochreiter 01] Hochreiter, S. and Bengio, Y., "Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies" (2001)

[Jia 20] Jia, Y., Min, G., Xu, C., Li, X., and Zhang, D., "A Knowledge Driven Dialogue Model With Reinforcement Learning", IEEE Access , Vol. 8, pp. 131741–131749 (2020)

[Johnson 17] Johnson, J., Douze, M., and Jégou, H., "Billion-scale similarity search with GPUs", arXiv preprint arXiv:1702.08734 (2017)

[Karpukhin 20] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t., "Dense Passage Retrieval for Open-Domain Question Answering", in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pp. 6769–6781, Online (2020), Association for Computational Linguistics

[Kim 21] Kim, B., Seo, S., Han, S., Erdenee, E., and Chang, B., "Distilling the Knowledge of Large-scale Generative Models into Retrieval Models for Efficient Open-domain Conversation", in Findings of the Association for Computational Linguistics: EMNLP 2021 , pp. 3357–3373, Punta Cana, Dominican Republic (2021), Association for Computational Linguistics

[Kwiatkowski 19] Kwiatkowski, T., Palomaki, J., Redfield, O., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S., "Natural Questions: A Benchmark for Question Answering Research", Transactions of the Association for Computational Linguistics , Vol. 7, pp. 453–466 (2019)

[Lakew 18] Lakew, S. M., Cettolo, M., and Federico, M., "A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation", in Proceedings of the 27th International Conference on Computational Linguistics , pp. 641–652, Santa Fe, New Mexico, USA (2018), Association for Computational Linguistics

[Lan 20] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R., "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", in International Conference on Learning Representations (2020)

[Le 19] Le, H., Sahoo, D., Chen, N., and Hoi, S., "Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems", in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pp. 5612–5623, Florence, Italy (2019), Association for Computational Linguistics

[Lee 23] Lee, H., Hou, R., Kim, J., Liang, D., Hwang, S. J., and Min, A., "A Study on Knowledge Distillation from Weak Teacher for Scaling Up Pre-trained Language Models", in Findings of the Association for Computational Linguistics: ACL 2023 , pp. 11239–11246, Toronto, Canada (2023), Association for Computational Linguistics

[Lewis 20a] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension", in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pp. 7871–7880, Online (2020), Association for Computational Linguistics

[Lewis 20b] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. eds., Advances in Neural Information Processing Systems , Vol. 33, pp. 9459–9474, Curran Associates, Inc. (2020)

[Li 16a] Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B., "A Diversity-Promoting Objective Function for Neural Conversation Models", in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pp. 110–119, San Diego, California (2016), Association for Computational Linguistics

[Li 16b] Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., and Dolan, B., "A Persona-Based Neural Conversation Model", in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 994–1003, Berlin, Germany (2016), Association for Computational Linguistics

[Li 17] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S., "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset", in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) , pp. 986–995, Taipei, Taiwan (2017), Asian Federation of Natural Language Processing

[Li 21] Li, L., Lin, Y., Ren, S., Li, P., Zhou, J., and Sun, X., "Dynamic Knowledge Distillation for Pre-trained Language Models", in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , pp. 379–389, Online and Punta Cana, Dominican Republic (2021), Association for Computational Linguistics

[Liu 16] Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J., "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation", in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing , pp. 2122–2132, Austin, Texas (2016), Association for Computational Linguistics

[Liu 20] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V., "Ro{BERT}a: A Robustly Optimized {BERT} Pretraining Approach" (2020)

[Melis 17] Melis, G., Dyer, C., and Blunsom, P., "On the state of the art of evaluation in neural language models", arXiv preprint arXiv:1707.05589 (2017)

[Oluwatobi 20] Oluwatobi, O. and Mueller, E., "DLGNet: A Transformer-based Model for Dialogue Response Generation", in Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI , pp. 54–62, Online (2020), Association for Computational Linguistics

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., "Bleu: a Method for Automatic Evaluation of Machine Translation", in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics , pp. 311–318, Philadelphia, Pennsylvania, USA (2002), Association for Computational Linguistics

[Raffel 20] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", J. Mach. Learn. Res. , Vol. 21, No. 1 (2020)

[Rashkin 19] Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L., "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset", in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pp. 5370–5381, Florence, Italy (2019), Association for Computational Linguistics

[Roller 21] Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., et al., "Recipes for Building an Open-Domain Chatbot", in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume , pp. 300–325 (2021)

[Sai 20] Sai, A. B., Mohankumar, A. K., Arora, S., and Khapra, M. M., "Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining", Transactions of the Association for Computational Linguistics , Vol. 8, p. 810–827 (2020)

[Sanh 19] Sanh, V., Debut, L., Chaumond, J., and Wolf, T., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", ArXiv , Vol. abs/1910.01108, (2019)

[Serban 17] Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y., "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues", in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence , AAAI'17, p. 3295–3301, AAAI Press (2017)

[Settles 08] Settles, B. and Craven, M., "An analysis of active learning strategies for sequence labeling tasks", in proceedings of the 2008 conference on empirical methods in natural language processing , pp. 1070–1079 (2008)

[Shao 17] Shao, Y., Gouws, S., Britz, D., Goldie, A., Strope, B., and Kurzweil, R., “Generating High-Quality and Informative Conversation Responses with Sequence-to-Sequence Models”, in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing , pp. 2210–2219, Copenhagen, Denmark (2017), Association for Computational Linguistics

[Shen 21] Shen, Z. and Xing, E. P., “A Fast Knowledge Distillation Framework for Visual Recognition”, in European Conference on Computer Vision (2021)

[Shuster 21] Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J., “Retrieval Augmentation Reduces Hallucination in Conversation”, in Findings of the Association for Computational Linguistics: EMNLP 2021 , pp. 3784–3803, Punta Cana, Dominican Republic (2021), Association for Computational Linguistics

[Sordoni 15] Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B., “A Neural Network Approach to Context-Sensitive Generation of Conversational Responses”, Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2015)

[Sutskever 14] Sutskever, I., Vinyals, O., and Le, Q. V., “Sequence to Sequence Learning with Neural Networks”, in Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 , NIPS’14, p. 3104–3112, Cambridge, MA, USA (2014), MIT Press

[Touvron 23] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G., “LLaMA: Open and Efficient Foundation Language Models” (2023)

[Tu 22] Tu, Q., Li, Y., Cui, J., Wang, B., Wen, J.-R., and Yan, R., “MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation”, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 308–319, Dublin, Ireland (2022), Association for Computational Linguistics

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I., “Attention is All You Need”, in Proceedings of the 31st International Conference on Neural Information Processing Systems , NIPS’17, p. 6000–6010, Red Hook, NY, USA (2017), Curran Associates Inc.

[Vinyals 15] Vinyals, O. and Le, Q. V., “A Neural Conversational Model”, in ICML Deep Learning Workshop (2015)

[Weizenbaum 66] Weizenbaum, J., “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine”, Commun. ACM , Vol. 9, No. 1, p. 36–45 (1966)

[Weston 18] Weston, J., Dinan, E., and Miller, A., “Retrieve and Refine: Improved Sequence Generation Models For Dialogue”, in Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd

International Workshop on Search-Oriented Conversational AI , pp. 87–92, Brussels, Belgium (2018), Association for Computational Linguistics

[Wolf 19] Wolf, T., Sanh, V., Chaumond, J., and Delangue, C., “TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents”, ArXiv , Vol. abs/1901.08149, (2019)

[Wolf 20] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, von P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M., “Transformers: State-of-the-Art Natural Language Processing”, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations , pp. 38–45, Online (2020), Association for Computational Linguistics

[Xu 22] Xu, Y., Ishii, E., Cahyawijaya, S., Liu, Z., Winata, G. I., Madotto, A., Su, D., and Fung, P., “Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters”, in Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering , pp. 93–107, Dublin, Ireland (2022), Association for Computational Linguistics

[Zhang 18] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J., “Personalizing Dialogue Agents: I have a dog, do you have pets too?”, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 2204–2213, Melbourne, Australia (2018), Association for Computational Linguistics

[Zhang 20a] Zhang, A., Chen, S., Zhang, X., Li, R., and Zhang, X., “A Knowledge-Enriched Model for Emotional Conversation Generation”, Companion Proceedings of the Web Conference 2020 (2020)

[Zhang 20b] Zhang, R., Zheng, Y., Shao, J., Mao, X., Xi, Y., and Huang, M., “Dialogue Distillation: Open-Domain Dialogue Augmentation Using Unpaired Data”, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pp. 3449–3460, Online (2020), Association for Computational Linguistics

[Zhou 18] Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B., “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory”, ArXiv , Vol. abs/1704.01074, (2018)

[Zhu 21] Zhu, Q., Chen, X., Wu, P., Liu, J., and Zhao, D., “Combining Curriculum Learning and Knowledge Distillation for Dialogue Generation”, in Findings of the Association for Computational Linguistics: EMNLP 2021 , pp. 1284–1295, Punta Cana, Dominican Republic (2021), Association for Computational Linguistics