

AMERICAN UNIVERSITY OF BEIRUT

METADIAL: A META-LEARNING APPROACH
FOR DIALOGUE GENERATION IN ARABIC
LANGUAGE

by

MOHSEN YOUSSEF SHAMAS

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Computer Science
of Faculty of Arts and Sciences
at the American University of Beirut

AMERICAN UNIVERSITY OF BEIRUT

METADIAL: A META-LEARNING APPROACH
FOR DIALOGUE GENERATION IN ARABIC
LANGUAGE

by
MOHSEN YOUSSEF SHAMAS

Approved by:

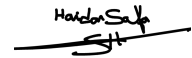
Dr. Wassim El Hajj, Professor
Computer Science

Advisor



Dr. Haidar Safa, Professor
Computer Science

Member of Committee



Dr. Shady Elbassuoni, Associate Professor
Computer Science

Member of Committee



Date of thesis defense: September 5, 2022

ACKNOWLEDGEMENTS

I am deeply thankful to my thesis advisor Professor Wassim El Hajj for his excellent guidance, extreme patience, and persistent help. Without him, this work would not have been feasible. He guided me through every step of this thesis and taught me how to do research and always encouraged me to do better.

My deepest gratitude must also be extended to the esteemed members of my thesis committee: Professor Shady Elbassuoni and Professor Haidar Safa; I would like to thank them for accepting to serve on my committee and helped me improve my work with their valuable comments. I am also extremely thankful to Professor Hazem Hajj and Professor Khaled Shaban for their valuable remarks throughout the research process. Special thanks to the U.S. Middle Eastern Partnership Initiative's scholarship program for funding my master's program. I would also like to express my gratitude to my University, the American University of Beirut for the amazing years where I got to meet and get to know the brightest professors and colleagues.

Finally, my deepest thanks go to my parents for their unconditional emotional support throughout these challenging two years of study. They always wished me the best in my studies and in my career.

ABSTRACT OF THE THESIS OF

Mohsen Youssef Shamas for Master of Science
Major: Computer Science

Title: MetaDial: A Meta-learning Approach for Dialogue Generation in Arabic Language

Dialogue generation is the automatic generation of a text response, given a post by a user. The advancements in deep learning models have made developing conversational systems not only possible, but also effective and helpful in many applications spanning a variety of domains. Nevertheless, work on Arabic Conversational bots is still limited due to various challenges including the language rich morphology, huge vocabulary, and the scarcity of data resources. Although meta-learning has been introduced before in the natural language processing (NLP) realm and showed significant improvements in many tasks, it has rarely been used in natural language generation (NLG) tasks and never in Arabic NLG. In this thesis, we propose a meta-learning approach for Arabic Dialogue generation for fast adaptation on low resource domains. We start by using existing pre-trained models; we then meta-learn the initial parameters on high resource dataset before fine-tuning the parameters on the target tasks. We prove that the proposed model that employs meta-learning techniques improves generalization and enables fast adaptation of the transformer model on low-resource NLG tasks. We report gains in the BLEU-4 in improvements in Semantic textual Similarity (STS) metrics in comparison with the existing state-of-the-art approach. We also do a further study on the effectiveness of the meta-learning algorithms on the response generation of the models.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	1
ABSTRACT	2
ABBREVIATIONS	8
1 Introduction	9
1.1 Motivation	9
1.2 Arabic Natural Language Processing	10
1.3 Objective	11
2 Literature Review	13
2.1 English Dialogue Generation	13
2.1.1 Meta-learning for Dialogue Generation	17
2.2 Arabic Dialogue Generation	17
3 Proposed Approach	20
3.1 Background: Meta-learning	20
3.2 Learning Framework	21

4	Datasets	26
4.1	ArabicTopicalChat Dataset	26
4.2	ArabicEmpatheticDialogue Dataset	29
4.3	Data Statistics	29
4.3.1	English Dataset Statistics	30
4.3.2	Arabic Dataset Statistics	30
4.3.3	Finetuning Language Models	31
5	Experiments and Results	32
5.1	Experimental Setup	32
5.2	Evaluation	32
5.2.1	Automatic Evaluation	33
5.2.2	Statistical Significance	35
5.2.3	Manual Evaluation	37
5.3	Discussion	38
5.4	Limitations	39
6	Conclusion and Future Work	41
6.1	Conclusion	41
6.2	Future Work	42
A	Distributions of Numerical Results of Manual Evaluation	43
B	Examples on Response Generation	47
	Bibliography	50

ILLUSTRATIONS

1.1	Example dialogue from ArabicTopicalChat (Translation of TopicalChat) between a human and an artificial chatbot	10
3.1	Meta-learning initial parameters	20
3.2	Encoder-Decoder Transformer Architecture	22
3.3	Meta-learning initial parameters	24
4.1	Good example translation from ArabicTopicalChat	28
4.2	Example from ArabicEmpatheticDialogue Dataset	28
5.1	Paired Bootstrap Resampling	36
A.1	Distribution of fluency scores in 10% domain by 3 models	44
A.2	Distribution of relevance scores in 10% domain by 3 models	44
A.3	Distribution of fluency scores in 30% domain by 3 models	45
A.4	Distribution of relevance scores in 30% domain by 3 models	45
A.5	Distribution of fluency scores in 50% domain by 3 models	46
A.6	Distribution of relevance scores in 50% domain by 3 models	46

B.1	Example 1 from generated responses	48
B.2	Example 2 from generated responses	48
B.3	Example 3 from generated responses	49

TABLES

4.1	Similar readability scores for the two datasets show that they are semantically and syntactically similar	29
4.2	Arabic translations of the datasets have similar statistics with ArabicEmpatheticDialogue scoring higher difficulty due to longer sentences and higher average Faseeh count	30
4.3	Evaluation of the finetuned transformers on different datasets using BLEU-3, BLEU-4 and Semantic Textual Similarity (STS)	31
5.1	automatic evaluation of the three approaches (FT: finetuned, MAML and finetuned, and reptile and finetuned) showing BLEU-4 and STS scores	34
5.2	automatic evaluation of the three approaches using the English datasets, showing BLEU-4 and STS scores	34
5.3	Paired bootstrap resampling performed on results generated by models trained on 30% training sub-dataset	36
5.4	Results of manual evaluation of the 9 experiments	37

ABBREVIATIONS

AI	Artificial Intelligence
AIML	Artificial Intelligence Markup Language
BERT	Bidirectional Encoder Representations from Transformers
GLUE	General Language Understanding Evaluation
GAN	Generative Adversarial Network
GPT	Generative Pretrained Transformer
GRU	Gated Recurrent Unit
MAML	Model-Agnostic Meta-Learning
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
RL	Reinforcement Learning
RNN	Recurrent Neural Network
Seq2Seq	Sequence-to-Sequence
SOTA	State-Of-The-Art

CHAPTER 1

INTRODUCTION

1.1 Motivation

Conversational chatbots are the artificial agents that perform the task of dialogue generation in either specific or open domains. Open domain chatbots are meant to enroll in broad range of general-topic human conversations. In classifying artificial chatbots based on the general approaches used to develop them, there have been generally two categories: the retrieval based chatbots and the generative chatbots. Retrieval based chatbots use repositories of predefined answers and heuristics to choose appropriate answers to input text. Generative chatbots, on the other hand, use machine translation techniques to generate response as a translation to the user input(context).

The development of dialogue generation systems started way before the development of the earliest generative deep learning models, where the first chatter bots [1], [2] relied purely on rule based approaches and primitive natural language pro-

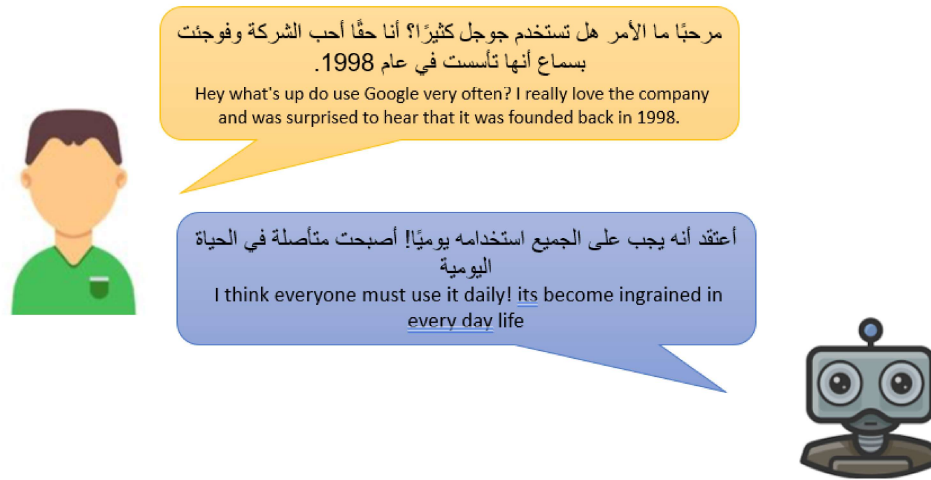


Figure 1.1: Example dialogue from ArabicTopicalChat (Translation of TopicalChat) between a human and an artificial chatbot

cessing scripts. However, with the development of powerful neural network models and their successful implementation on many natural language processing (NLP) and natural language generation (NLG) applications, the development and use of dialogue response systems became ubiquitous. Artificial conversational agents are currently being used in a wide variety of applications from mental health care systems, to IT support services, to marketing, to entertainment industry and many others. Focus on open-domain chatbots started to arise after the development of complex recurrent neural networks(RNNs) and lately transformers architectures to study NLG tasks of which lies the dialogue response generation.

1.2 Arabic Natural Language Processing

Although many approaches have been used to develop robust conversational systems, such systems can impose a real challenge when it comes to the Arabic language. Arabic is in itself a sophisticated language with a complex morphology and rich

vocabulary. This contributes heavily to the difficulty of developing Arabic-speaking artificial agents. In addition to the complex characteristics of the language itself, the availability of data for training or evaluation is still very limited, in comparison to other high resource languages. The challenge of data scarcity in the Arabic NLP tasks has been studied and pursued in previous works such as utilizing semi-supervised learning for Arabic Named Entity Recognition [3]. Given those challenges and others, there has yet to be a powerful and efficient model for developing a robust Arabic dialogue system.

1.3 Objective

This work introduces meta-learning to the realm of Arabic natural language generation, and specifically Arabic dialogue generation. We provide a comprehensive study on the effectiveness of using meta-learning techniques on pretrained transformers for Arabic dialogue generation. To achieve our goal, we adopt the full transformer architecture [4] and initialize it with pretrained checkpoints from the AraBERT model [5]. To mitigate the challenges of low resources, and to further enable the fast adaptation of the model on small datasets, we employ the famous Model Agnostic Meta-Learning (MAML) algorithm [6] and its variation, Reptile [7]. The aim of this optimization approach is to enable fast adaptation of the model when finetuned on small datasets[6]. This approach has been shown to be successful when trained on classification tasks from the GLUE benchmark [8] especially when Reptile algorithm was used to meta-learn initial parameters[9]. To the best of our

knowledge, there has not been any work on Arabic NLG that studies the effect of meta-learning algorithms on training Arabic NLP models. Also, our approach is the first to apply meta-learning on the full transformer model for the open-domain dialogue generation task.

The rest of the work is organized as follows: in chapter 2, we review the literature on dialogue generation in English and Arabic languages. Next, we define our approach for tackling Arabic dialogue generation in chapter 3. Then, datasets are described in chapter 4. We proceed to discuss our experiments and results in chapter 5, and we provide conclusion and final remarks in chapter 6.

CHAPTER 2

LITERATURE REVIEW

Researchers have developed plenty of approaches to improve the quality of response generation by the conversational agents. In this chapter, we will first focus on the notable English state of the art approaches and then review the attempts to tackle Arabic Dialogue generation.

2.1 English Dialogue Generation

Before the advancements in deep learning approaches, chatbots used to utilize markup languages and knowledge bases and databases as retrieval based models. [10] developed a corpus based conversational agent that utilizes SQL to retrieve responses from database. They developed a real time system, which could extract both database attributes and attribute values from the user input and automatically respond to the user's input using a rule based approach. [11] studied and reviewed reward-driven process, partially observable Markov decision processes (POMDPs). In their work, they explored POMDP-based spoken dialog systems and explained

the challenges and mitigations such reward based dialogue systems. An example of an end-to-end system was proposed by [12]. Their work studied practical and efficient end-to-end dialog control with supervised and reinforcement learning. Their model, Hybrid Code Networks, combines a recurrent neural network with knowledge templates to reduce the amount of training data required and to benefit from inferring a latent representation of dialog state. Rasa is an open-source intent-based chatbot framework [13]. The framework is composed of different components: speech recognition module, spoken language understanding (SLU) module, dialogue state tracker, dialogue policy optimization, and a Text to Speech module. [14] use statistical machine translation (SMT) approaches and information retrieval (IR) approaches to develop dialogue response generation models. They show in their work that SMT-based models perform substantially better than IR-based models. With the development of the sequence to sequence architectures [15], that are comprised of two recurrent neural networks, named encoder and decoder, chatbots would have the capability of response generation. [16] developed an ensemble of retrieval-based and generation-based dialogue system. The response is generated by a recurrent neural network. [17] developed a sequential matching network that first matches response with each utterance in the context and creates a vector with matching information. Vectors are fed into an RNN that outputs the response. [18] explored the issue of generating safe and commonplace responses encountered by researchers when developing sequence-to-sequence models; they found that one reason behind this is the use of unidi-rectional likelihood of output (responses) given input (messages). They propose Maximum Mutual Information (MMI) as

the objective function for the Seq2Seq model to mitigate this problem and showed that their approach outperform previous approaches with respect to the diversity of the generated outputs. In yet another approach, the work of [19] presented two persona-based neural response generation models; a single-speaker SPEAKER MODEL and a dyadic SPEAKER-ADDRESSEE MODEL, within a sequence-to-sequence (SEQ2SEQ) framework. They reported improvements in the BLEU and perplexity scores after incorporating persona to their dialogue response system. In other attempts to tackle dialogue generation, some researchers have used reinforcement learning (RL) and Generative Adversarial Networks (GANs) to produce and study dialogue response models; [20] proposed using an adversarial training approach for response generation and trained their model in a reinforcement learning framework. They report performance improvements when training their model using adversarial learning. [21] introduced MILABOT, a deep reinforcement learning chatbot that consists of an ensemble of natural language generation and retrieval models, including template-based models, bag-of-words models, sequence-to-sequence neural network and latent variable neural network models. [22] modeled dialogue generation as a reinforcement learning problem where they defined a reward function for a Seq2Seq model based on Gated Recurrent Units (GRU) and attention. In an attempt to utilize transformers for developing conversational bots, [23] introduced a new benchmark EmpatheticDialogue, and used it to train and evaluate deep learning models. They employed workers from Amazon Mechanical Turk to generate the conversations and the empathy labels for each context-response pair. They experimented with different variations of the transformer model (auto-encoder, auto-

regressive decoder, and full architecture). They observed that the produced response by the created chatbots contain higher level of empathy than previous approaches. In another approach tackling dialogue generation in the few-shot domain, [24] explored dialogue generation using pretrained language model. Their work introduces and defines the problem of few-shot learning in NLG. It also proposed a multidomain table to text dataset. And they propose a novel algorithm to reduce human annotation efforts and improve model performance. [25] developed an improved chatbot that uses EmpatheticDialogue dataset to finetune a generative pretrained transformer (GPT) pretrained on the BooksCorpus dataset [26], a big dataset that contains more than 7000 unpublished books. Also, they pretrained the model on the PersonaChat dataset [27] to increase engagingness of the model. They showed that pretraining a transformer model yields better generalization and improvement in its ability of natural language understanding (NLU). Other approaches modeled the Dialogue generation task as reinforcement learning problem in which the chatbot interacts with the end-users and observes the results of its actions. It receives each time a reward which can be positive or negative. The learning process of the chatbot happens throughout the conversations. [28] presented LaMDA: Language Models for Dialog Applications. The work describes a family of Transformer-based neural language models specialized for dialog that, when finetuned on annotated data, lead to significant improvements of the dialogue response generation towards the two key challenges of safety and factual grounding. They also explore the use of LaMDA in the multiple domains including education and content-recommendations.

2.1.1 *Meta-learning for Dialogue Generation*

Recent work on dialogue generation has started utilizing meta-learning technique to optimize the natural language generation (NLG) performed by the conversational agent; [29] proposed a domain adaptive dialog generation method based on meta-learning (DAML). Their model, DAML, learns from multiple rich-resource tasks and then adapts to new domains with minimal training samples. The two-step gradient updates in DAML enable the model to learn general features across multiple tasks. They observed that DAML proves as robust and effective method for training dialogue systems with low resources. In a similar approach, [30] proposed a generalized optimization-based meta-learning approach Meta-NLG for the low-resource NLG task. Meta-NLG utilizes Meta NLG tasks and a meta-learning optimization procedure based on Model Agnostic Meta-Learning (MAML). They showed in their work that the Meta NLG outperforms all existing models on the low resource tasks that they tested on and proved that their approach fastens the adaptation significantly in low resource situations. In a more recent approach, [31] proposed a Meta- X_{NLG} framework based on meta-learning and language clustering for effective cross-lingual transfer and generation. The work studies the use of meta-learning for zero-shot cross-lingual transfer and generation.

2.2 Arabic Dialogue Generation

Although research on conversational systems has been myriad in general, contributions in Arabic dialogue generation and chatbots development is still limited. This

is due to the challenges imposed by the language structure and sophisticated morphology, and also due to the limited data for benchmarking and evaluation. It was in 2014 that the first work tackling Arabic conversational bot was published ArabChat [32]. The architecture was comprised a scripting engine and a rule-based scripting language to handle the contexts of conversations. The authors also developed a Pattern Matching approach to handle users' conversations. ArabChat was able to direct the user in their conversation's topics, and at the same time it was flexible enough to follow the user and their topics of interest. [33] introduced "BOTTA", an Arabic Chatbot that simulated conversations in the Egyptian Dialect. AIML was used to develop the chatbot, and AIML files were used to create BOTTA's knowledge base. Moreover, sets of AIML from ROSIE (variation of ALICE chatbot [2]) were directly translated to be used by BOTTA. OlloBot [34] is an Arabic conversational agent that aims to assist physicians support patients with the health process. The development of OlloBot was done by listing intents and entities and building the dialogue structure and flow on IBM Watson Conversation. More recently, work on Arabic Chatbots started to involve more deep learning methods. [35] introduced an empathetic chatbot adopting a deep learning approach; the authors trained a Bi-LSTM Seq2Seq model combined with attention on the ArabicEmpatheticDialogue data set. Their contribution showed that deep learning yielded better results than all previous approaches. Another work that uses the same dataset [36] experimented with the transformer model [4] with pretrained checkpoints and outperformed the previous model when evaluated using the same data from ArabicEmpatheticDialogue.

There has been some approaches in tackling dialogue generation and natural

language generation in Arabic and English, and meta-learning has been studied for the NLG task, but the work is still limited to NLG in English, and end-to-end task oriented dialogue generation. Our approach is the first to study the effect of meta-learning on Arabic dialogue generation, and the first to study open-domain dialogue generation for low resource language using the full sequence-to-sequence transformer architecture with meta-learning a language model.

CHAPTER 3

PROPOSED APPROACH

In this chapter, we first provide background on meta-learning algorithms, then we elaborate on incorporating meta-learning with the deep learning framework to achieve fast adaptation for dialogue generation.

3.1 Background: Meta-learning

Meta-learning is the process of learning to learn. Meta-learning algorithms are optimization algorithms that aim to optimize specific part of the machine learning

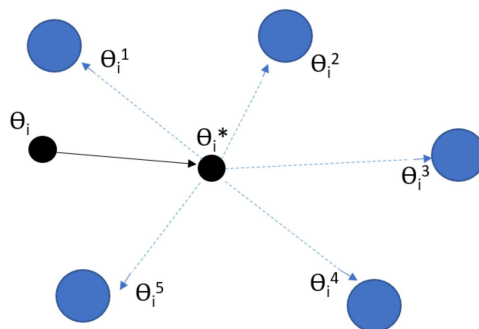


Figure 3.1: Meta-learning initial parameters

process. Some algorithms aim to learn to initialize model parameters[6], others focus on learning the optimizer throughout the learning process[37]; some other algorithms learn to compare[38], and others aim to learn the whole learning process [39]. In our work we incorporate MAML with the learning framework; MAML is an algorithm that aims to initialize the model parameters to fastly adapt to the target tasks at the finetuning stage, using as little data as possible. The goal of MAML optimization is to enable the model of adapting to low resource target tasks by meta-training it on high-resource auxiliary tasks. Formally, We define a set of tasks $T = \{T_1, T_2, T_3, \dots\}$ given that task $T = \langle X, Y \rangle$ where $X = x_1x_2\dots x_m$ and $Y = y_1y_2\dots y_n$ are the respective sequences of words that comprise input and response sentences. Model parameters θ are meta-learned using auxiliary tasks T_{aux} and then finetuned on target tasks:

$$\theta^* = Learn(T_t, Metalearn(T_{aux}, \theta))$$

where θ^* are the parameters adapted to the low resource task T_t .

The hypothesis tested in our work proposes that using MAML for Arabic dialogue generation would initialize the model parameters to be closer to the domain space of the target tasks(see figure 3.1).

3.2 Learning Framework

Our framework is comprised of three main phases: pretraining of a transformer model, meta-learning the parameters of the pretrained model using the meta-learning

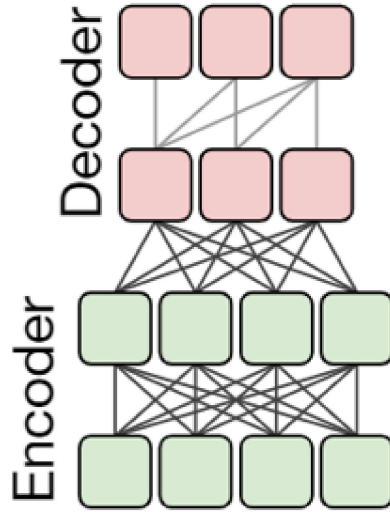


Figure 3.2: Encoder-Decoder Transformer Architecture

algorithm, and lastly finetuning the model on the target task. The full framework is illustrated in figure 3.3.

The model architecture that we adapt extends the BERT2BERT model architecture developed by [40]. It is essentially a seq2seq model that uses the original transformer implementation of the work of [4] with slight modifications. For the encoder implementation, the RELU activation function is replaced with the GELU function following the BERT model [41]. The decoder implementation is also identical to BERT but with 1 modifications: the self attention mechanism is masked to look only to the left context. An attention mechanism is added between the encoder and the decoder. Illustration of the model architecture can be shown in figure 3.2

We load a pretrained model with both encoder and decoder parameters initialized with pretrained checkpoints. Pretraining is the unsupervised learning of the model parameters on unlabeled NLP tasks to increase generalization [41]. For the pre-finetuning optimization, meta-learning, we experiment with model-agnostic meta-

learning [6] which, uses auxiliary tasks to meta-learn model’s initial parameters to enable fast adaptation of the model on target tasks. The method is gradient descent-based, as it samples a batch of tasks from a task distribution and performs multiple training steps. Given the initial parameters of the model and the list of the updated parameters over each training step, the next step is to learn parameters that can adapt to each of the tasks in a balanced fashion.

In the context of the NLP task of dialogue generation, a task is defined as an input-response data sample. the meta-learning phase is done by sampling auxiliary tasks from a task distribution in batches of constant size k , and for each batch the model learn k parameters from the tasks. Then, the meta-learning algorithm performs an outer update to the model parameters from the k updated parameters. Formally, given model f_θ parametrized by θ and distribution $p(T)$ over set of tasks T_1, T_2, \dots, T_k , at each iteration of the meta-learning algorithm, we sample batch of tasks $\{T_i\}$ of size k from $p(T)$. Then for each task T_i , model parameters are updated with k gradient descent steps using the equation:

$$\theta_i^{(k)} = \theta_i^{(k-1)} - \alpha \nabla_{\theta^{k-1}} L_i(f_{\theta_i^{k-1}})$$

where α is the meta-learning rate and L_i is the objective function chosen to be the categorical cross-entropy error over the task T_i .

The meta-learning step following the k computations of the parameters $\theta_i^{(k)}$ would then differ between the original MAML algorithm and Reptile algorithm. This steps is used to update the original parameters θ .

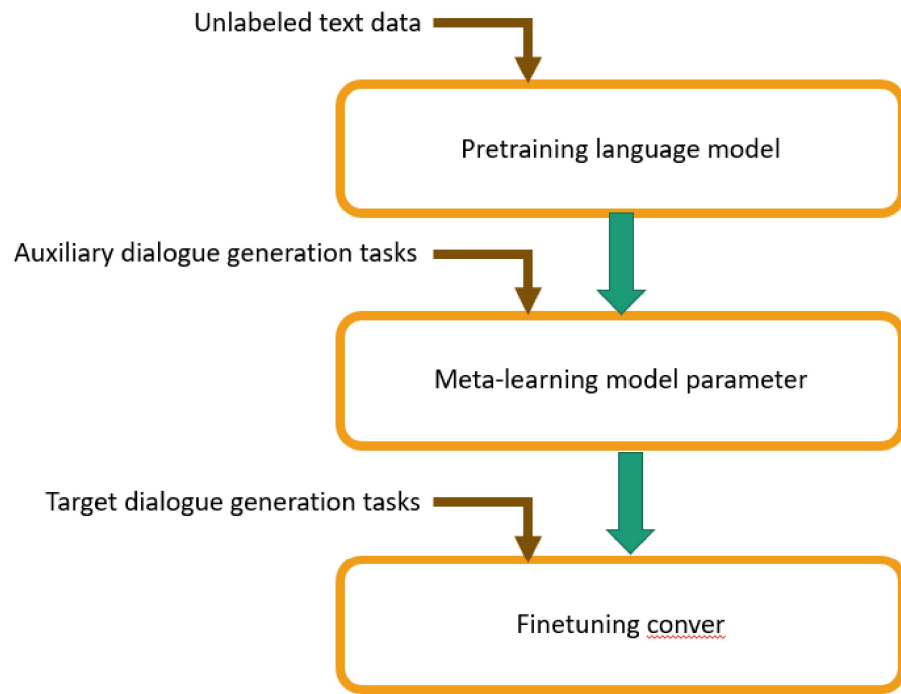


Figure 3.3: Meta-learning initial parameters

The vanilla MAML algorithm updates the model with meta-objective function[9]:

$$\min \sum_{T(i) \sim p(T)} L(f_{\theta_i^k})$$

hence the parameters θ are updated using MAML by using gradient descent:

$$\theta = \theta - \beta \nabla_{\theta} L(f_{\theta_i^k})$$

where β is the meta-learning rate of f_{θ} over θ .

On the other hand, reptile algorithm does not use gradient update for a second time; it uses the multiple gradient descents from the previous steps to move the model weights toward new parameters. The equation by which the parameters are updated according to reptile is:

$$\theta = \theta + \beta \frac{1}{\|T_i\|} \sum_{T(i) \sim p(T)} (\theta_i^k - \theta)$$

Although reptile algorithm has a much lower complexity than the vanilla MAML algorithm because MAML performs second order gradient descent which is computationally heavy and time consuming, it has been demonstrated in previous works [7], [9], [42] that reptile can achieve competitive and sometimes better performance.

CHAPTER 4

DATASETS

Almost every Arabic Natural Language Processing project encounters the challenge of securing datasets that fairly represents the NLP application’s data space. Native datasets for Arabic dialogue generation are still non-existent, so we use Arabic translations from existing English datasets for the task of dialogue response generation. In this chapter, we first describe the dataset of which the auxiliary tasks are sampled for meta-learning the model parameters. Then, we describe the dataset used of which the target tasks are used for finetuning the model for dialogue generation. We also compare the datasets against each other in terms of readability, and by finetuning language models using each of them.

4.1 ArabicTopicalChat Dataset

We sample the auxiliary tasks used for meta-learning from ArabicTopicalChat, a dataset we created by translating the TopicalChat dataset [43] from English to

Arabic using the python library `googletrans`¹ that implements the google translate API. The original dataset contains 11 thousands human-human conversation, where each conversation is alternating dialogue lines between two human conversing agents. The total number of conversation lines is approximately 188 thousands, distributed over 8 sentiment classes.

To create ArabicTopicalChat, we translate every conversation line from TopicalChat and then for each two consecutive lines that belong to the same conversation, we create one data point with the first line as context and the second line as response. This is done for the purpose of meta-learning using tasks similar in shape to the target tasks. After that, we clean the dataset from sentences with remaining English words, as they would affect the meta-learning negatively. ArabicTopicalChat dataset contains 123025 context-response pair of Arabic human conversation lines. Examples from the dataset can be shown in the figures 1.1 and 4.1

We then evaluate quality of the translation of the dataset following [35]. We sample 100 random sentences from TopicalChat dataset with their respective translations and compare each Arabic sentence to its Arabic translation. Assessing the quality of the translation, we find out that 6 out of the 100 sample translations were unreasonable whereas 95 were reasonable. This was due to several factors one of which is the informality of the conversations and sometimes the slang phrases used by the English speakers that cannot be properly translated to Arabic.

¹<https://pypi.org/project/googletrans/>

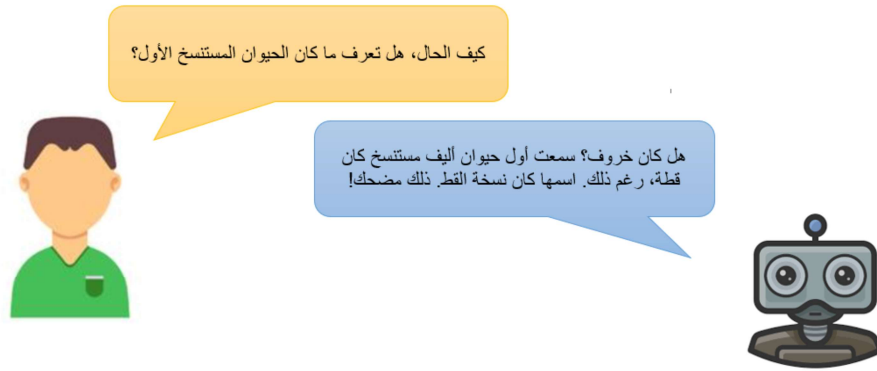


Figure 4.1: Good example translation from ArabicTopicalChat



Figure 4.2: Example from ArabicEmpatheticDialogue Dataset

Metrics	EmpatheticDialogue	TopicalChat
Flesch	85.43	86.71
Cunning Fog	6.53	6.12
Dale-Chall	7.62	6.49
Automated Readability Index	4.6	3.09

Table 4.1: Similar readability scores for the two datasets show that they are semantically and syntactically similar

4.2 ArabicEmpatheticDialogue Dataset

ArabicEmpatheticDialogue Dataset was introduced in the work of [35], and was used to train the first deep learning model for Arabic dialogue generation. It was also used for finetuning the existing state-of-the-art model [36]. Hence, following the recent successful work, we use ArabicEmpatheticDialogue to finetune our model. The target tasks set size is the same size of the training dataset described in the literature. ArabicEmpatheticDialogue contains 36626 context-response pair of Arabic human conversation lines; the conversations are distributed over 6 sentiment classes. The conversations were translated from English EmpatheticDialogue[23] also using the Googletrans API. An example from the dataset is presented in figure 4.2.

4.3 Data Statistics

Statistics performed on the two English datasets and two Arabic datasets and compared to each others respectively show that both datasets, Arabic and English, contain semantically and syntactically similar data.

Metrics	ArabicEmpatheticDialogue	ArabicTopicalChat
Average Faseeh count	0.17	0.02
Average Char count	72.7	53.8
Average Sentence length	9.98	7.78

Table 4.2: Arabic translations of the datasets have similar statistics with ArabicEmpatheticDialogue scoring higher difficulty due to longer sentences and higher average Faseeh count

4.3.1 *English Dataset Statistics*

We evaluated English datasets using Flesch, Cuning Fogg, automated readability index (ARI) [44], and Dave-Chall[45] metrics. Each of these metrics calculates a readability score that indicate a certain level of the difficulty for reading English language text. Similar scores for Flesch and Cuning Fog signifies that both datasets are readable by people from the same educational level; high scores for Flesch, Fog and A.R.I.(between 80 and 90 for Flesch and between 6 and 7 for Fog) signifies lower difficulty levels which means that text from both datasets is readable by middle school students. The results are shown in table 4.1.

4.3.2 *Arabic Dataset Statistics*

Since evaluation of Arabic text is more challenging, we performed statistics on the word count and character count per conversation line in addition to the faseeh count[46] which counts the number of "faseeh" words, which are words in Arabic language with morphological and structural aspects that, if exist in a sentence, increase the difficulty of reading it. Scores for the Arabic datasets are shown in table 4.2. The statistics indicate that ArabicTopicalChat dataset is slightly less complex and and more general than ArabicEmpatheticDialogue dataset.

Metrics\Training data	ArabicEmpatheticDialogue	ArabicTopicalChat cleaned
BLEU-3	0.1028	0.1174
BLEU-4	0.09	0.1207
STS	0.5912	0.522

Table 4.3: Evaluation of the finetuned transformers on different datasets using BLEU-3, BLEU-4 and Semantic Textual Similarity (STS)

4.3.3 *Finetuning Language Models*

To demonstrate the efficacy of the newly introduced dataset ArabicEmpatheticDialogue dataset, we finetune a language model using each dataset. We use the BERT2BERT transformer architecture and initialize the model with AraBERT pre-trained checkpoints following [35]. We evaluate the models using the testing split from ArabicEmpatheticDialogue and report the BLEU-3, BLEU-4, and the semantic textual similarity scores. The results are shown in the table 4.3.

CHAPTER 5

EXPERIMENTS AND RESULTS

5.1 Experimental Setup

We initialize a Seq2Seq transformer with AraBERT checkpoints. We meta-learn the model using auxiliary tasks from ArabicTopicalChat; we learn on 6000 task batches randomly sampled from the dataset with a batch size of 32 tasks. Huggingface’s Transformers implementation of the transformer architecture was used [47]. Pytorch library code was utilized to implement MAML and Reptile, huggingface trainer API was used for finetuning the models on ArabicEmpatheticDialogue dataset, and 12GB NVIDIA Tesla K80 GPU was used for accelerating the meta-learning and finetuning processes.

5.2 Evaluation

We train and evaluate three different models: pretrained model meta-learned using MAML (Pretrained+MAML+Finetuned), pretrained model meta-learning us-

ing reptile (Pretrained+Reptile+Finetuned), and the existing state-of-the-art model [35] which does not use any form of optimization (Pretrained + Finetuned). Each of the models is finetuned on different sized subsets of the original ArabicEmpatheticDialogue dataset to interpret the efficiency of MAML and Reptile optimization on different sized training data: 10%, 30%, and 50%.

We generate the text response using top-k sampling with $k=50$, as opposed to beam search which was used by the most recent work [36]. It has been shown that using top-k sampling with large values of k would decrease the repetitiveness of the model output and makes it more human-like[48].

5.2.1 *Automatic Evaluation*

We use the Bilingual Evaluation Understudy (BLEU) automated metric [49] and the semantic textual similarity (STS) metric to evaluate our experiments. BLEU is a metric mainly introduced for evaluating automatic machine translation models. It works by comparing the generated output with the reference "translations", and computes the overlap between n-grams from generated text and the reference. The score range is between zero and hundred with zero indicating no overlap at all, and one meaning that the reference and the candidate are identical. The semantic textual similarity metric compares the reference response to the model-generated response by computing the sentence encoding of each response and computing the cosine similarity between them.

Models	10% data		30% data		50% data	
	BLEU-4	STS	BLEU-4	STS	BLEU-4	STS
FT	6.82	64.82	8.91	68.68	8.63	67.73
MAML.FT	7.62	67.27	9.17	68.75	9.01	68.16
Reptile.FT	7.91	67.12	9.19	68.73	8.95	67.85

Table 5.1: automatic evaluation of the three approaches (FT: finetuned, MAML and finetuned, and reptile and finetuned) showing BLEU-4 and STS scores

Models	10%		30%		50%	
	BLEU-4	STS	BLEU-4	STS	BLEU-4	STS
FT	8.1601	24.7291	9.8139	28.387	10.5662	30.891
MAML + FT	8.2956	23.8921	9.777	28.7763	10.9723	31.3401
Reptile + FT	8.2796	24.3395	9.5993	27.9365	10.7232	30.9173

Table 5.2: automatic evaluation of the three approaches using the English datasets, showing BLEU-4 and STS scores

Arabic Dialogue Systems Automatic Evaluation

We evaluate our 3 models, after training each one on 3 different-sized training datasets separately, resulting in 9 experiments. We report BLEU-4 and STS scores in table 5.1. Our results show that both MAML and Reptile score higher for STS and BLEU-4 when trained on a very small dataset. However, as the size of the dataset increases, BLEU and STS scores for the three approaches converges to nearly equal results. One conclusion that can be concluded from our results is that meta-learning is most efficient in the low-resource domains with the scarcest resources.

English Dialogue Systems Automatic Evaluation

To get further insights on our results, we use the original English datasets to conduct the same experiments. We meta-learn our models using MAML and reptile algorithms. We use EmpatheticDialogue dataset for meta-learning. For finetuning

we use 10%, 30%, and 50% subsets from TopicalChat dataset. We also evaluate the models using BLEU and STS. Results are shown in table 5.2. Our results show that both MAML and Reptile score higher than the existing state-of-the-art model for STS and BLEU-4 when trained on very small datasets. The tables also show that BLEU scores for all English experiments is larger than all BLEU scores for Arabic experiments given the same size of the training data.

5.2.2 *Statistical Significance*

To get more insights on how significant is the improvements in the BLEU scores between the existing state-of-the-art model and our meta-learned models. Hence, we choose to test the statistical significance of the models finetuned on 30% of the training data as it presents the average medium among our mediums of experiments. We adopt the paired bootstrap resampling test as described in [50]. Bootstrap and resampling are widely applicable statistical methods which relax many of the assumptions of classical statistics. Bootstrap allows computation of statistics from limited data and allows us to compute statistics from multiple subsamples of the dataset. It also allows us to make minimal distribution assumptions.

Therefore, in order to test the statistical significance of our BLEU results on our models outputs, we sample subsets from the 1800 results generated by two systems, each subset has a size half to the population size which is 900, and compare the BLEU-4 scores. We repeat $k=200$ times with different samples. After that, we use the pairs of the computed BLEU scores to compute the p-value, win-lose ratio of the models, median scores, mean scores, and 95% confidence intervals. The overall

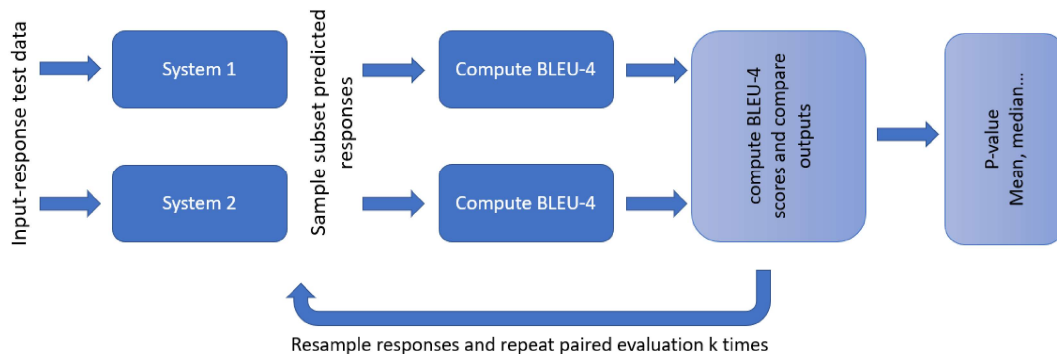


Figure 5.1: Paired Bootstrap Resampling

system pairs	models	win-ratio	mean	median	CI	p
Rep vs FT	Rep	0.870	0.139	0.139	[0.134, 0.144]	0.130
	FT	0.130	0.136	0.139	[0.131, 0.142]	-
MAML vs FT	MAML	0.540	0.152	0.153	[0.133, 0.177]	0.460
	FT	0.460	0.152	0.152	[0.127, 0.172]	-
MAML vs Rep	MAML	0.020	0.133	0.133	[0.129, 0.139]	-
	Rep	0.980	0.139	0.139	[0.135, 0.145]	0.020

Table 5.3: Paired bootstrap resampling performed on results generated by models trained on 30% training sub-dataset

procedure is illustrated in figure 5.1.

Our results show that there seems to be a trend showing that the model meta-learned with Reptile performs better than the model that was meta-learned as the p-value is equal to 0.13 which signifies that the probability that the null hypothesis (the hypothesis that the improvement in the scores is sheer coincidence) is false is 87%. On the other hand the we can reject the null hypothesis that MAML does not improve over the state-of-the-art with only 54% confidence. Numerical details are provided in the table 5.3. We note that the confidence interval in all experiments is narrow so we can conclude that the sampled data is sufficient for generalizing over the data space.

Models	10%		30%		50%	
	Fluency	Relevance	Fluency	Relevance	Fluency	Relevance
FT	2.59	1.83	3.40	2.12	3.4	2.43
MAML + FT	3.18	2.7	3.54	2.77	3.75	2.87
Reptile + FT	2.69	2.06	3.42	2.78	3.62	2.92

Table 5.4: Results of manual evaluation of the 9 experiments

5.2.3 *Manual Evaluation*

It has been shown that BLEU metric is insufficient alone for evaluating conversational systems. [51] suggests that automated metrics used for evaluating machine translation and automatic text summarization tasks are insufficient and weakly correlate with human evaluation. For this reason, and because of the lack of an established standardized evaluation methodology, we perform manual evaluation for our models. We choose to optimize the 3 models finetuned using the 3 subsets from the training dataset (10%, 30%, and 50%). So the total number of experiments evaluated manually is 9 experiments.

We generate the text response using top-k sampling with k=50, as opposed to beam search which was used by the most recent work naous-etal-2021-empathetic. It has been shown that using top-k sampling with large values of k would decrease the repetitiveness of the model output and makes it more human-like[48]. We sample 50 examples from the testing dataset from ArabicEmpatheticDialogue. For each of the 50 examples, each one of the 9 models would generate a response, given input context. The generated responses are then rated by speakers of the Arabic language (one evaluation to each example). The evaluator is asked to rate the sentences fluency and its relevance to the given context. To do so, the rater is asked to answer

two questions regarding each sentence’s fluency and relevance:

- Fluency: How understandable was the generated response from a language perspective
- Relevance: How relevant was the generated response to the given input context?

Each sentence is given 2 ratings, each ranges from 1 to 5, with 1 scoring the lowest, and 5 being the highest rating for any sentence. The evaluation results are shown in table 5.4. More on the distribution of the manual evaluation results for each experiments are provided in Appendix A. Observing the ratings from different experiments, the histograms show a clear shift in ratings between the previous SOTA model and the meta-learned models as the meta-learned models generated responses scored more times 5 and 4 than the previous SOTA model in all experiments, and the ratings for the meta-learned models’ generated output scored higher, according to table 5.4.

5.3 Discussion

To get more insights on the output, we sample some examples and generate the responses using the English and Arabic models trained on 30% of the data. The examples are provided in Appendix B.

Observing the generated output from the models, it can be deduced that the meta-learning algorithms are improving the models’ performances from several perspectives. In addition to increasing coherence and fluency and improving relevance

of the generated output, the meta-learned models seem to show tendency to predict sentiment and emotional state from the input better than the non-meta-learned models. This occurs even when the models do not produce completely coherent output. Also, the English meta-learned models seem to produce fluent outputs no matter how little data the models are being trained on. This does not stand for Arabic experiments, as we observe that some generated responses contain spelling mistakes, and sometimes the whole generated sentence lacks coherence. This also endorses the difference in the BLEU scores, as English BLEU scores were higher than the Arabic scores in general, as stated before. This can be attributed to the error margin and the weakness in the translation of the training and testing data, and also to the complex morphology and huge vocabulary, in comparison to the English language. This can be further explored from a linguistic perspective to understand the difference between NLG in English and Arabic and to point out distinguishing attributes of each of the two languages.

5.4 Limitations

Limitations of the work span different areas. A major limitation imposed on us and on researchers in the domain of Arabic Natural Language Generation, is the non-existence of any native public dataset for the Arabic NLG, which imposes a huge challenge on developing robust Arabic dialogue system. Another limitation is the computational power and complexity limitation, which is specific to meta-learning. since the algorithm is sequential by construction, even using GPUs with high VRAM

for parallel processing does not accelerate the training enough. Consequently, the meta-learning process can be inefficiently time-costly especially if the domain space is relatively huge which requires substantial amount of data to be used for training.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this section, we summarize the key points from this thesis and state the future work to be done.

6.1 Conclusion

In this work, we present a novel comprehensive study on optimizing Natural Language Generation models in Arabic language using two meta-learning algorithms: vanilla model-agnostic meta-learning and Reptile. To evaluate our proposed learning framework, we introduce a dataset for meta-learning, ArabicTopicalChat. We evaluate our models using BLEU and STS automated metrics and test the significance of the improvement of the BLEU scores with respect to previous approach by employing paired bootstrap resampling. We further demonstrate that meta-learning natural language generation models produce improved response generation by manually evaluate sample generated responses. We hence showed that in the Arabic low-resource domain, meta-learning can improve results of dialogue generation.

6.2 Future Work

This research opens the doors for many works that can still be done regarding studying meta-learning for Arabic NLG and Arabic NLP in general. Firstly, the work can be improved by working on producing a native Arabic dataset for open-domain dialogue response generation, as this will eliminate the challenge of the translation error margin completely. Furthermore, more extensive experimental study on different sized datasets can yield more robust conclusions regarding most efficient task set sizes. Another way that we believe that can substantially improve the results is tuning the hyperparameters for meta-learning. Experiments can be conducted more extensively to study the effect of sampling more or less batches, and also the effect of increasing or decreasing the size of the sampled batches from auxiliary tasks. Another plan for the future experiments is to experiment with multi-domain and task-oriented dialogue generation datasets in English and Arabic. Other future trajectories include experimenting with different datasets from low resource domains (low resource languages, domain specific dialogue generation...) and with other meta-learning algorithms.

APPENDIX A

DISTRIBUTIONS OF NUMERICAL RESULTS OF MANUAL EVALUATION

While it shows clearly that the meta-learned models outperform the state-of-the-art, we can notice that the improvement is still margin and the increase trend is tangential to the previous approach. Also, it can be established that both reptile and MAML improve the generation of the NLG model, but more experiments should be conducted to enable the researchers of drawing conclusions regarding the efficiency of MAML algorithm with respect to Reptile, and when should Reptile be used, and when should MAML be used on the other hand. This is among the important future directions that this work may continue in to explore distinguishing characteristics between MAML and Reptile.

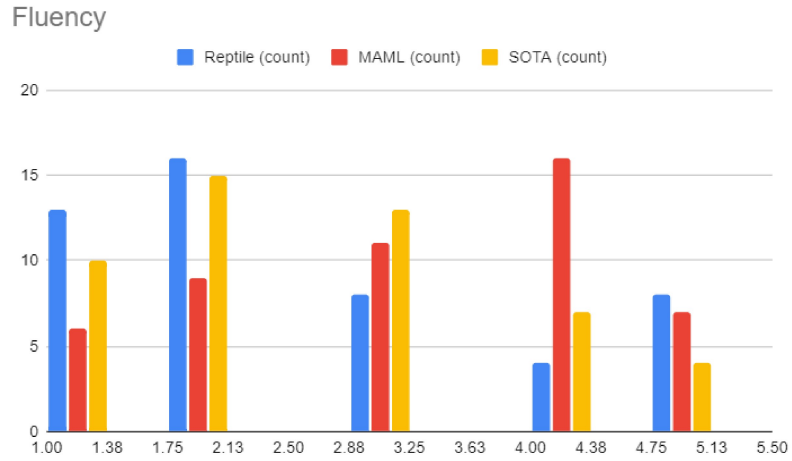


Figure A.1: Distribution of fluency scores in 10% domain by 3 models

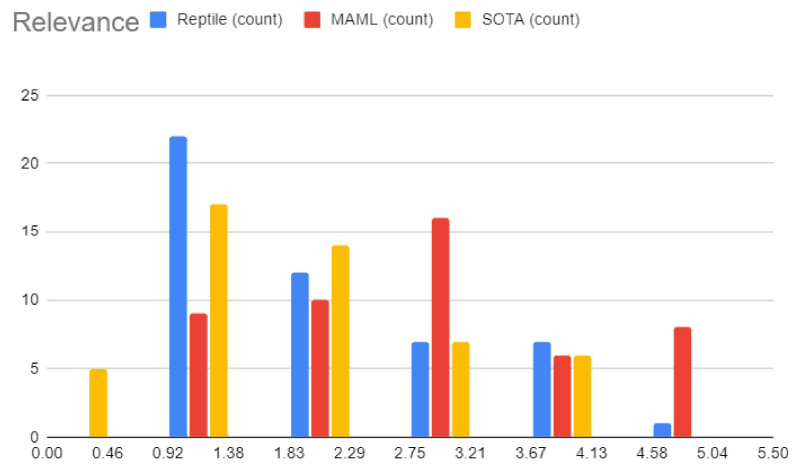


Figure A.2: Distribution of relevance scores in 10% domain by 3 models

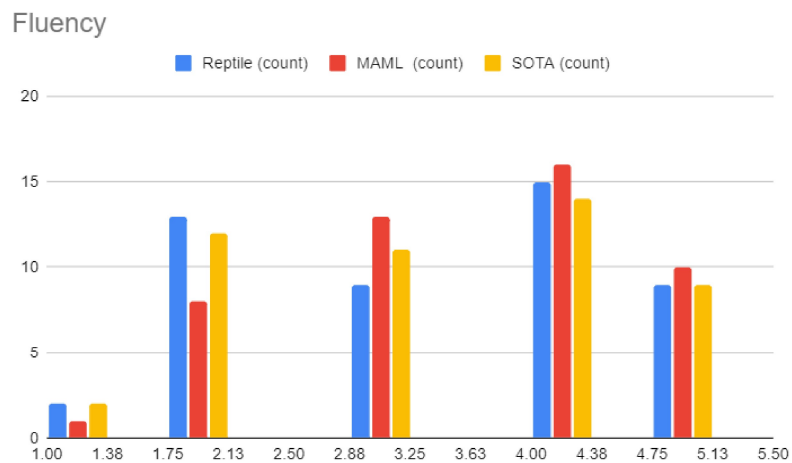


Figure A.3: Distribution of fluency scores in 30% domain by 3 models

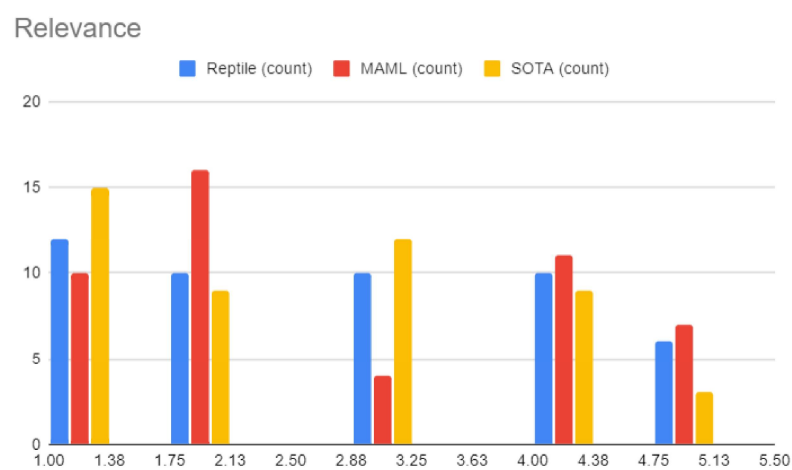


Figure A.4: Distribution of relevance scores in 30% domain by 3 models

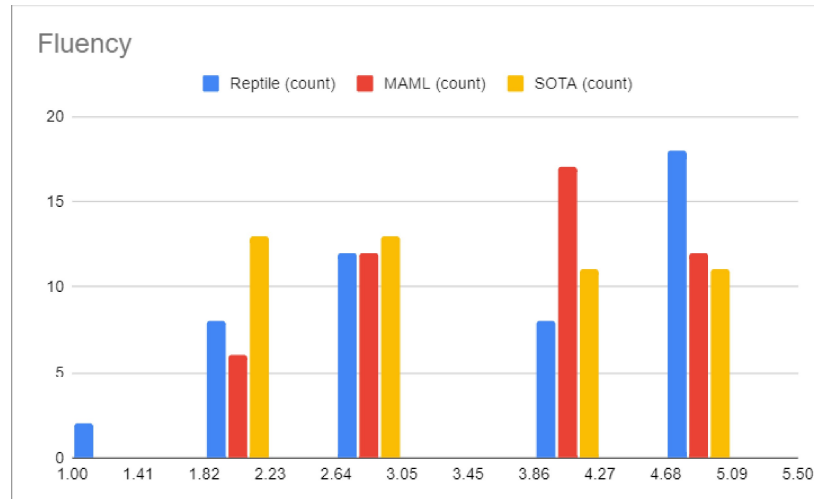


Figure A.5: Distribution of fluency scores in 50% domain by 3 models

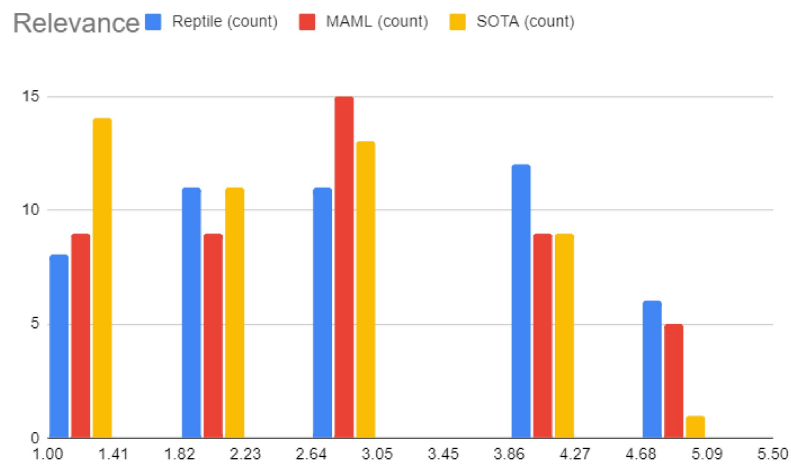


Figure A.6: Distribution of relevance scores in 50% domain by 3 models

APPENDIX B

EXAMPLES ON RESPONSE GENERATION

We provide examples from training MAML, Reptile and SOTA models using 30% of the training data from the English and the Arabic domain in this appendix.

Arabic	Context	لقد كان مقفلا . كسروا النافذة وسرقوا كل ما عندي من نقود وهدية عيد ميلاد ابني التي كانت في صندوق السيارة .
	Reference response	هل تمزح معي؟ العصب من بعض الناس. ما مدى اليأس الذي يجب أن تكون عليه... حسنا ، أمل أن تتمكن على الأقل من استبدال الحاضر حتى لا يكون ابنك حزينا للغاية. وعلى الأقل لم يحاولوا اقتحام منزلك وإيذائك.
	MAML response	هل حصلت على تعويض مالي ؟
	Reptile response	هذا مروع ، كان عليك دفع دولار . يجب أن تكون ، مثل سيارة أخرى للإظيف ولقاع باسمه سيارتك في رحلتك
	Finetuned response	هذا مروع ، كان عليك دفع دولار .!
English	Context	It was locked. They broke the window and stole all my change and my son's birthday present that was in the trunk.
	Reference response	Are you kidding me? The nerve of some people. How desperate do you have to be... Well, I hope you can at least replace the present so your son isn't too sad. And at least they didn't try to bust into your home and harm you.
	MAML response	how do you feel about it now?
	Reptile response	that stinks. Im sorry you lost it. would that be a big thing to do?
	Finetuned response	i stole my car and my brother stole the cash from me. I felt very alone the entire time.

Figure B.1: Example 1 from generated responses

Arabic	Context	قابلت صديقا قديما في المدرسة الثانوية أثناء تواجدي في المتجر . كانت صدمة !
	Reference response	هل مضى وقت طويل؟
	MAML response	أنت محظوظ.. ماذا حدث ؟
	Reptile response	أنا أعرف كيف يصبح ذلك .
	Finetuned response	لماذا ؟ لماذا ؟ هل شعرت لماذا تعتقد أنك فعلت ذلك مرة أخرى ؟
English	Context	I met an old friend in high school while I was at the store. It was a shock!
	Reference response	Has it been a long time?
	MAML response	thats great! how was it?
	Reptile response	Yes, the whole way back to school was <u>pretty cool</u> .
	Finetuned response	thats awesome. did you two meet in high school?

Figure B.2: Example 2 from generated responses

Arabic	context	أنا مستاءة للغاية . اكتشفت للتو أن زوجي كان يخونني الأسبوع الماضي .
	Reference response	أوه ، هذا مقرف . هل تحدثتم يا رفاق أو قطعته على الفور؟
	MAML response	آسف لسماع ذلك . ماذا حدث لك ؟
	Reptile response	هل كانوا بخير ؟ اتمنى انه لا بد من أن يكون الأمر
	Finetuned response	لا يوجد شيء سيء ، هل كان على الطعام للصرغم من ذلك ؟
English	context	I am very upset. I just found out my husband was cheating on me last week.
	Reference response	Oh, this is disgusting. Did you guys talk or cut it straight away?
	MAML response	oh no! thats terrible, how did he manage to lose it?
	Reptile response	i lost my husband and he had to lie about it.
	Finetuned response	oh no! are you really upset about?

Figure B.3: Example 3 from generated responses

BIBLIOGRAPHY

- [1] J. Weizenbaum, “Eliza—a computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
- [2] R. Wallace, “Artificial linguistic internet computer entity (alice),” *City*, 1995.
- [3] C. Helwe, G. Dib, M. Shamas, and S. Elbassuoni, “A semi-supervised BERT approach for Arabic named entity recognition,” in *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 49–57. [Online]. Available: <https://aclanthology.org/2020.wanlp-1.5>.
- [4] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [5] W. Antoun, F. Baly, and H. Hajj, “Arabert: Transformer-based model for arabic language understanding,” *arXiv preprint arXiv:2003.00104*, 2020.
- [6] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*, PMLR, 2017, pp. 1126–1135.
- [7] A. Nichol, J. Achiam, and J. Schulman, “On first-order meta-learning algorithms,” *arXiv preprint arXiv:1803.02999*, 2018.
- [8] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “Glue: A multi-task benchmark and analysis platform for natural language understanding,” *arXiv preprint arXiv:1804.07461*, 2018.
- [9] Z.-Y. Dou, K. Yu, and A. Anastasopoulos, “Investigating meta-learning algorithms for low-resource natural language understanding tasks,” *arXiv preprint arXiv:1908.10423*, 2019.
- [10] K. Pudner, K. A. Crockett, and Z. Bandar, “An intelligent conversational agent approach to extracting queries from natural language,” in *World Congress on Engineering*, vol. 1, 2007, p. 305.
- [11] S. Young, M. Gašić, B. Thomson, and J. D. Williams, “Pomdp-based statistical spoken dialog systems: A review,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

- [12] J. D. Williams, K. Asadi, and G. Zweig, “Hybrid code networks: Practical and efficient end-to-end dialog control with supervised and reinforcement learning,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 665–677. DOI: [10.18653/v1/P17-1062](https://doi.org/10.18653/v1/P17-1062). [Online]. Available: <https://aclanthology.org/P17-1062>.
- [13] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, “Rasa: Open source language understanding and dialogue management,” *arXiv preprint arXiv:1712.05181*, 2017.
- [14] A. Ritter, C. Cherry, and W. B. Dolan, “Data-driven response generation in social media,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 583–593. [Online]. Available: <https://aclanthology.org/D11-1054>.
- [15] K. Cho, B. Van Merriënboer, C. Gulcehre, *et al.*, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [16] Y. Song, R. Yan, X. Li, D. Zhao, and M. Zhang, “Two are better than one: An ensemble of retrieval-and generation-based dialog systems,” *arXiv preprint arXiv:1610.07149*, 2016.
- [17] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li, “Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots,” *arXiv preprint arXiv:1612.01627*, 2016.
- [18] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, “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*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 110–119. DOI: [10.18653/v1/N16-1014](https://doi.org/10.18653/v1/N16-1014). [Online]. Available: <https://aclanthology.org/N16-1014>.
- [19] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, “A persona-based neural conversation model,” *arXiv preprint arXiv:1603.06155*, 2016.
- [20] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, “Adversarial learning for neural dialogue generation,” *arXiv preprint arXiv:1701.06547*, 2017.
- [21] I. V. Serban, C. Sankar, M. Germain, *et al.*, “A deep reinforcement learning chatbot,” *arXiv preprint arXiv:1709.02349*, 2017.
- [22] J. Shin, P. Xu, A. Madotto, and P. Fung, “Generating empathetic responses by looking ahead the user’s sentiment,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7989–7993.

- [23] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, “Towards empathetic open-domain conversation models: A new benchmark and dataset,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5370–5381. DOI: [10.18653/v1/P19-1534](https://doi.org/10.18653/v1/P19-1534). [Online]. Available: <https://aclanthology.org/P19-1534>.
- [24] Z. Chen, H. Eavani, W. Chen, Y. Liu, and W. Y. Wang, “Few-shot nlg with pre-trained language model,” *arXiv preprint arXiv:1904.09521*, 2019.
- [25] Z. Lin, P. Xu, G. I. Winata, *et al.*, “Caire: An end-to-end empathetic chatbot,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 622–13 623.
- [26] Y. Zhu, R. Kiros, R. Zemel, *et al.*, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.
- [27] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, “Personalizing dialogue agents: I have a dog, do you have pets too?” *arXiv preprint arXiv:1801.07243*, 2018.
- [28] R. Thoppilan, D. De Freitas, J. Hall, *et al.*, “Lamda: Language models for dialog applications,” *arXiv preprint arXiv:2201.08239*, 2022.
- [29] K. Qian and Z. Yu, “Domain adaptive dialog generation via meta learning,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2639–2649. DOI: [10.18653/v1/P19-1253](https://doi.org/10.18653/v1/P19-1253). [Online]. Available: <https://aclanthology.org/P19-1253>.
- [30] F. Mi, M. Huang, J. Zhang, and B. Faltings, “Meta-learning for low-resource natural language generation in task-oriented dialogue systems,” *arXiv preprint arXiv:1905.05644*, 2019.
- [31] K. Maurya and M. Desarkar, “Meta-x_{NLG}: A meta-learning approach based on language clustering for zero-shot cross-lingual transfer and generation,” in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 269–284. DOI: [10.18653/v1/2022.findings-acl.24](https://doi.org/10.18653/v1/2022.findings-acl.24). [Online]. Available: <https://aclanthology.org/2022.findings-acl.24>.
- [32] M. Hijjawi, Z. Bandar, K. Crockett, and D. Mclean, “Arabchat: An arabic conversational agent,” in *2014 6th International Conference on Computer Science and Information Technology (CSIT)*, 2014, pp. 227–237. DOI: [10.1109/CSIT.2014.6806005](https://doi.org/10.1109/CSIT.2014.6806005).
- [33] D. Abu Ali and N. Habash, “Botta: An Arabic dialect chatbot,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 208–212. [Online]. Available: <https://aclanthology.org/C16-2044>.

- [34] A. Fadhil and A. AbuRa'ed, "OloBot - towards a text-based Arabic health conversational agent: Evaluation and results," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 295–303. DOI: [10.26615/978-954-452-056-4_034](https://doi.org/10.26615/978-954-452-056-4_034). [Online]. Available: <https://aclanthology.org/R19-1034>.
- [35] T. Naous, C. Hokayem, and H. Hajj, "Empathy-driven Arabic conversational chatbot," in *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 58–68. [Online]. Available: <https://aclanthology.org/2020.wanlp-1.6>.
- [36] T. Naous, W. Antoun, R. Mahmoud, and H. Hajj, "Empathetic BERT2BERT conversational model: Learning Arabic language generation with little data," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 164–172. [Online]. Available: <https://aclanthology.org/2021.wanlp-1.17>.
- [37] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.
- [38] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.
- [39] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, PMLR, 2016, pp. 1842–1850.
- [40] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 264–280, 2020.
- [41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [42] Y. Tian and P. J. Gorinski, "Improving end-to-end speech-to-intent classification with reptile," *arXiv preprint arXiv:2008.01994*, 2020.
- [43] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, *et al.*, "Topical-chat: Towards knowledge-grounded open-domain conversations.," in *INTERSPEECH*, 2019, pp. 1891–1895.
- [44] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.
- [45] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948.

- [46] M. El-Haj and P. E. Rayson, “Osman: A novel arabic readability metric,” 2016.
- [47] T. Wolf, L. Debut, V. Sanh, *et al.*, “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.
- [48] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” *arXiv preprint arXiv:1904.09751*, 2019.
- [49] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [50] P. Koehn, “Statistical significance tests for machine translation evaluation,” in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 388–395.
- [51] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, “How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation,” *arXiv preprint arXiv:1603.08023*, 2016.