AMERICAN UNIVERSITY OF BEIRUT

FEW-SHOT LEARNING FOR CONVERSATIONAL BOTS IN LOW-RESOURCE SETTINGS

by
TAREK NABIL NAOUS

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Electrical and Computer Engineering
of the Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
March 2022

AMERICAN UNIVERSITY OF BEIRUT

FEW-SHOT LEARNING FOR CONVERSATIONAL BOTS IN LOW-RESOURCE SETTINGS

by

TAREK NABIL NAOUS

Approved by:

Dr. Hazem Hajj, Associate Professor Electrical and Computer Engineering	Advisor
Dr. Rabih Jabr, Professor Electrical and Computer Engineering	Member of Committee
Dr. Wassim El Hajj, Professor Computer Science	Member of Committee
Dr. Khaled Shaban, Professor Computer Science and Engineering, Qatar University	Member of Committee Khaled Shaban

Date of thesis defense: March 17, 2022

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Naous (Last) Tarek (First) Nabil (Middle)

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes

___ **As of the date of submission of my thesis**

✓ **After 1 year from the date of submission of my thesis .**

___ **After 2 years from the date of submission of my thesis .**

___ **After 3 years from the date of submission of my thesis .**

TarekNaous — Signature

April 6, 2022 — Date

ACKNOWLEDGMENTS

I begin this section of my thesis by thanking The Almighty Creator, The Lord of Mercy, my Sustainer and my Guardian. I owe my Lord Allah for giving me the motivation and determination during difficult times, the favor of good health, the blessing of worldly knowledge, and placing the people in my life who helped me complete my thesis. Any contribution I make is certainly imperfect, it is only He who is Perfect, beyond what any mind can imagine.

I thank my wonderful supervisor Prof. Hazem Hajj, who is one the nicest people I have met. It has been a pleasant journey working under his supervision. I have learned a lot from him both on a technical and professional level, helping me grow as a researcher and as a person. He has always been very encouraging and supportive, as if I was his own son. I will always remember him as an influential person in my journey towards becoming a world-class researcher. Although we did not meet in person due to the pandemic and the difficult circumstances in the country at this time, I hope we get the chance to meet in the future.

This acknowledgement would no go without thanking my wonderful lab mates at the AUB MIND Lab, some of whom have become friends and collaborators. My thanks also goes to my research supervisors Prof. Wassim El Hajj and Prof. Khaled Shaban. They have both been a joy to interact with and receive feedback from, pushing the quality of my work to a higher level. I also thank Prof. Rabih Jabr for agreeing to be on my thesis committee. I am grateful to have met such wonderful people.

I thank my parents and my brothers for their love and encouragement. My father has been the greatest example for me; sacrificing his time to provide good quality education for me and my brothers. Hard work, dedication, and honesty in work are some of the few things I have learned from him on an ethical level. I am blessed to have the love of a wonderful mother, who also sacrificed her time to raise her children. My brothers will always be the joy of my heart.

Finally, I would like to thank the American University of Beirut and the Electrical and Computer Engineering Department for providing me with the environment to thrive as a scholar, and a generous graduate fellowship that aided me in completing my master's degree. I will always look back at my time at AUB as a joyful experience full of wonderful memories. AUB will always be a home of mine.

ABSTRACT
OF THE THESIS OF

Tarek Nabil Naous for Master of Engineering
Major: Electrical and Computer Engineering

Title: Few-shot Learning for Conversational Bots in Low-Resource Settings

Open-domain dialogue agents are systems that can converse with users on any topic of user's choice. Having such types of agents has been a long standing objective in Artificial Intelligence as they can make the human-computer interaction experience much more engaging. Recent advances in English open-domain dialogue have leveraged state-of-the-art Large Language Models (LLMs) for Natural Language Generation (NLG). Such LLMs are massively pre-trained on unlabeled data in a self-supervised mode to learn abstract representations of the language. They also require large amounts of labeled open-domain dialogue data for fine-tuning to achieve the challenging task of dialogue response generation. In low-resource settings such as Arabic and its dialects, such pre-trained LLMs and large labeled dialogue datasets are often non-existent, hindering the development of open-domain chatbots for those languages. Such limited resource modeling problem is known as the few-shot learning problem.

In this thesis, we address multiple aspects of the few-shot learning problem for open-domain Arabic conversational bots. The first contribution is a solution to overcome the unavailability of LLMs with large amounts of labeled dialogue data for Arabic MSA. To address the response generation problem, we propose a model that transfers knowledge from a pre-trained BERT encoder to an encoder-decoder model for dialogue response generation. The second contribution addresses a more extreme case of limited resources with Arabic dialects. To address the LLM and NLG challenges for Arabic dialects, we propose a three-stage learning framework based on warm-starting, self-supervised pre-training, and few-shot fine-tuning. The third contribution focuses on addressing the challenge of ensuring generated responses are relevant to user's query for both English and Arabic. We propose a new decoding algorithm that considers increased samples in response generation then chooses the response with

highest similarity to user's query. The fourth contribution is in the development of new data resources for Arabic with one message-response dataset in Modern Standard Arabic (MSA) and three datasets for the most widely spoken Arabic dialects (Levantine, Egyptian, and Gulf). The experiment results showed success of the proposed methods and achieved state of the art performance for Arabic open-dialogue systems.

TABLE OF CONTENTS

ILLUSTRATIONS

TABLES

CHAPTER 1

INTRODUCTION

Open-domain conversational models aim to seamlessly blend knowledge and intelligence while satisfying users' need for communication and social belonging [1]. A long-standing goal of Artificial Intelligence (AI) has been to build intelligent open-domain conversational models that can understand the semantics of input utterances and provide coherent and relevant responses [2].

An important aspect of developing human-like chatbot models is enabling their sense of empathy. Empathy is described as the ability of recognizing others' state of mind and making sense of their feelings such as acknowledging others' pain, showing interest, gratitude, being supportive, or providing encouragement [3], [4]. Empathy is an innate capacity in most human beings, and is also described as a responsive and spontaneous act of copying an implied feeling. It triggers a sense of concern for others, leading to appropriate emotional reactions that instills a positive effect on interacting individuals. For instance, empathetic behavior is applicable to situations .

Conversational models with empathetic responding capabilities are crucial in making human-machine interactions closer to human-human interactions, as they can lead to increased engagement, more trust, and reduced frustration [5]. An important factor towards developing human-like chatbots is enabling their empathetic capability [6]. These characteristics are particularly desirable in open-domain conversational models as they can boost user satisfaction and make chatbots look less boorish. To this end, there has been a significant interest in developing empathetic conversational models [1], [2], [7], [8], where the models infer the emotions of a human user and provide a suitable empathetic response. The desired behavior is illustrated in Figure 1.1, where the empathetic agent recognizes that the user is feeling proud and, thus, generates an empathetic response that congratulates the user with enthusiasm.

Recent work on open-domain empathetic conversational models have adopted neural-based sequence generation approaches [9]. These approaches are based on encoder-decoder neural network architectures such as Sequence-to-Sequence (Seq2Seq) recurrent neural network models [10], or transformers [11]. The English literature also benefits from pre-trained language generation models and pre-trained conversational datasets that help in transferring knowledge and achieving high performance [12].

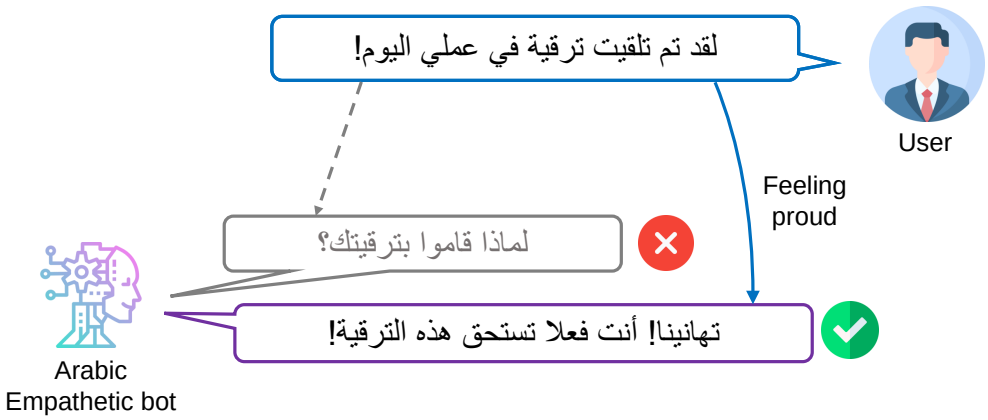

Figure 1.1: Example of empathetic behavior in an Arabic open-domain chatbot.

Despite the existence of valuable resources to build open-domain conversational models, most of them are in English, making it challenging to produce similar models for other languages, especially for languages that have many dialects. The low-resource challenge has been previously studied in the literature for task-oriented conversational models [13], machine translation [14], part-of-speech tagging [15], question-answering [16], and other NLP applications [17]. However, very little work targeted the issue of low-resources in open-domain conversational models.

Specifically for Arabic and its dialects, research on Arabic conversational models is still in its infancy mainly due to the lack of resources in terms of datasets and pre-trained models. Despite the availability of Arabic pre-trained language models such as AraBERT [18], which have proven useful for Arabic NLU tasks, the lack of pre-trained models for Arabic NLG makes the development of neural-based Arabic conversational models a challenging task. Hence, existing works on Arabic chatbots have mainly focused on retrieval-based methods [19] or rule-based approaches [20], [21]. While these approaches work well on task-oriented objectives, they are limited by the size of manually crafted rules they follow or the richness of the database they can retrieve responses from. This makes it difficult for such types of models to operate well in open-domain conversational settings, where generative neural-based models would be more suitable.

The aim of this thesis is to address low-resource challenges hindering the development of neural open-empathetic domain dialogue models in Modern Standard Arabic and Arabic dialects. The key research challenges to be addressed can be summarized as follows:

1. **Limited Data Samples:** There exist no previous datasets for open-domain dialog in both MSA or in dialects, which makes it difficult to learn dialogue response generation models.

2. **Learning Semantic and Syntactic Information of the Language:** Given the lack of good pre-trained models for Arabic dialects, it will be extremely chal-

lenging for the learning algorithm to capture the language semantic and syntactic information, which enables it to capture the meaning in the user's input.

3. **Learning Response Generation Skills:** Given the lack of conversational data to work with it is challenging for the model to learn how to provide a topic-relevant response from the small datasets it is trained on.

4. **Learning to Provide Empathetic Responses:** The bot is required to develop an empathetic behaviour. Hence, it needs to be able to recognize the emotion in the user's utterance and reply with an emotionally appropriate and topic-relevant response.

5. **Arabic-specific Challenges:** A main challenge specific to the Arabic language is the existence of multiple dialects with no standard orthography to follow and limited to no dialect-specific resources.

6. **Decoding Technique:** There exists several decoding algorithms to select from for sequence generation. Deterministic approaches results in repetitive and boring responses while probabilistic approaches provide more richness at the risk of potential increased irrelevancy in the output.

The objective of this thesis is to develop models, learning strategies, and decoding techniques for open-domain empathetic response generation in the Arabic language, including both MSA and DA. We adopt a Seq2Seq generative approach for response generation which has shown promise in the English literature for open-domain empathetic chatbots. However, different approaches to train the model need to be followed depending on the nature of Arabic that the model needs to deal with (MSA or DA), the size of the labeled dataset that it will be trained on, and the availability of resources that can be used for knowledge transfer. In what follows, an overview of the components used to develop our models are described.

The components that summarize the main methods proposed in this thesis are illustrated in Fig. 1.2. The main block is a Seq2Seq model that is made up of an encoder and a decoder. The encoder part is responsible for Natural Language Understanding (NLU), which enables the bot to understand the topic in the input utterance and recognize the emotion of the user. For Natural Language Generation (NLG), the decoder part uses the encoding provided by the encoder to generate a topic-relevant response in Arabic while exhibiting an empathetic behavior.

In order for the Seq2Seq model to perform well, it needs to be trained on a dataset of open-domain empathetic message-response pairs. Hence, the second important component of the system is the Dataset Creation part. When considering MSA, a large enough dataset (e.g., 40K samples) can be translated from an existing dataset in English using automatic translation tools such as the Google Translate API. Hence, given the translated labeled dataset, we can use it to perform regular supervised learning of the Seq2Seq model. This would result in a model that can generate open-domain responses in MSA while showing empathetic behavior.

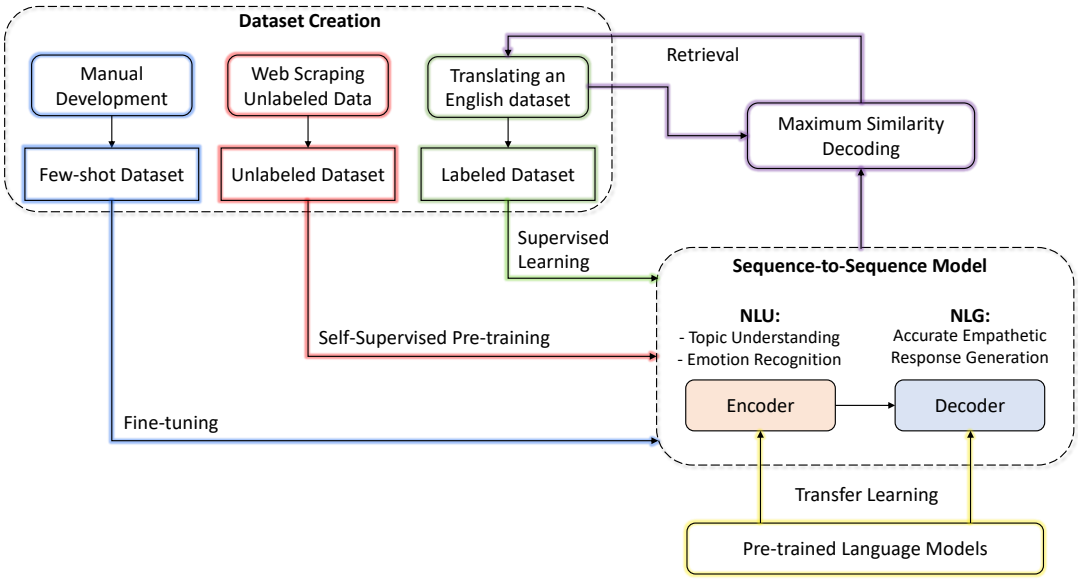

Figure 1.2: Diagrammatic view of the approaches proposed in this thesis

Training the model from scratch on the labeled dataset may not be enough to obtain very good results since the model would need to learn semantic and syntactic information of the language and the skill to generate topic-relevant and empathetic responses from the provided dataset, which may be relatively small in size. To overcome this issue, we perform transfer learning by initializing the encoder and decoder parts of the Seq2Seq model by the parameters of a pre-trained language model. This step transfers semantic and syntactic information of the language to Seq2Seq model so that it does not have to learn it from scratch. The model can then be fine-tuned in a supervised learning manner on the labeled dataset to learn the skill of response generation. By leveraging prior knowledge of the language, the model will perform better compared with from-scratch training on the labeled dataset.

The above mentioned approaches would work well for datasets in MSA. However, if the bot were expected to operate using a specific Arabic dialect, different approaches would need to be adopted both on the level of dataset creation and model training. First, no accurate tools for translation to DA are available which prevents us from directly translating an English dataset to DA. Second, pre-trained language models available in Arabic have been pre-dominantly trained on corpora in MSA, hence they do not perform very well on tasks in DA. To overcome these issues, we resort to two additional approaches in the dataset creation module. First, a dataset of utterance-response pairs is manually developed by writing equivalent samples from an English dataset. Since this procedure is manually done, this would result in a few-shot dataset that would have a very small size (e.g., 1K samples) and would not be useful for training a Seq2Seq model from scratch. Second, an unlabeled dataset of sentences

in DA can be obtained by web-scraping social media platforms such as Twitter, where most of the content generated by users is written in DA. The unlabeled dataset obtained can be fairly large in size (e.g., 1 million samples) since the scraping process is done in an automated manner. Thus, this dataset can be used to pre-train the Seq2Seq model in a self-supervised manner, helping the model acquire prior knowledge of the language semantic and syntactic information. After self-supervised pre-training, the model is then fine-tuned on the few-shot dataset to learn response generation in DA.

Finally, a decoding technique needs to be selected to generate the response from the model. Existing deterministic approaches such as greedy search or beam search result in highly repetitive and "boring" responses. On the other hands, using introducing randomness through probabilistic approaches such as top-k or top-p sampling with temperature would provide more rich and exciting responses. This, however, raises an additional risk of going off-topic. The selection of the hyper-parameters such as p and t are is also an ad-hoc process. We provide a way to automatically select those parameters and enhance relevance through a maximum similarity decoder which leverages knowledge from the training data using a retrieval module.

The rest of this report is organized as follows: In Chapter 2, we first provide background on the problem in learning from little data, define some learning paradigms that address this issue, and review existing methods to solve this problem. In Chapter 3, the first work on open-domain empathetic dialogue response generation in MSA is presented, providing a description on the dataset creation process and results achieved after training a baseline model. In Chapter 4, we address the problems faced in the earlier work where poor performance was achieved by adopting a transfer learning approach for sequence generation using warm-started transformers. In Chapter 5, we propose a learning framework based on three stages of warm-starting, self supervised pre-training, and fine-tuning to learn open-domain dialog models in low-resource Arabic dialects. In Chapter 6, we address the problems observed with the usage of existing decoding techniques and propose a new algorithm for decoding that is targeted at enhancing response relevance in open domain dialog. Concluding remarks and future directions are discussed in Chapter 7.

CHAPTER 2

LEARNING FROM LITTLE DATA: BACKGROUND

A key aspect of human intelligence is the ability to establish their cognition to novelty only from a few examples. While this task is easy for humans to do, it is very challenging for machine learning algorithms. Supervised learning methods require thousands of labeled data samples to achieve generalization. Achieving generalization from a handful of data samples with supervised information remains a challenge for machine learning models. In this regard, there exist many learning approaches beyond vanilla supervised learning to handle the problem of few samples. Many real world problems can benefit from those learning methods, such as learning for rare cases, learning for low-resource languages, in addition to the reduction of data collection and annotation costs [22].

2.1 Learning Paradigms and Definitions

We define some learning paradigms that are used to address the problem of learning with little data. It is noted that in the literature, some of those learning paradigms are used interchangeably.

- **Few-shot Learning:** Few-shot learning is a learning paradigm that has been proposed to tackle this challenge [23]. Few-shot learning methods leverage **prior knowledge** to generalize from the few labeled samples available. Prior knowledge is any information the learning algorithm has about the unknown hypothesis before being trained on the few data samples. Prior knowledge may be in the form of datasets from related tasks, pre-trained models, etc.

- **Transfer Learning:** Transfer learning leverages information learned from a source task to perform better on a related target task by providing good parameter initialization. Transfer learning can be used to solve certain few-shot learning problems [24], [25].

- **Meta Learning:** Meta learning approaches leverage meta-knowledge extracted by a meta-learner across several tasks to improve performance on a task-specific dataset [26]. Meta-learning, as will be discussed in a later section, has been used an an approach to solve few-shot learning problems where the task-specific dataset only has a few examples.

- **Weakly-supervised Learning:** Weakly-supervised learning methods deal with classification or regression problems where data samples are incomplete, noisy, or inaccurate. Semi-supervised learning is a weakly-supervised learning approach where a small number of data samples have supervised information and a much larger amount of data samples are unlabeled [27].

2.2 Few-shot Learning in Computer Vision

2.2.1 *Multi-task Learning*

In multi-task learning, models are trained to perform multiple related tasks at once, making them suitable for adoption in few-shot learning scenarios. Consider a source task for image classification of several species of monkey, where a dataset is available containing a large number of labeled images for each of the species. On the other hand, consider the target task of classifying rare species of monkeys for which only a few labeled images are available. Multi-task models learn both task-generic and task-specific information by joint training on both tasks. Multi-task learning through parameter sharing is a strategy, shown in Fig. 2.1, that shares parameters across the tasks by using shared layers at the beginning to learn common information, and then task-specific layers at the end to deal with different outputs for each task [28]. By leveraging prior knowledge from related datasets at the level of the shared layers, performance on classifying samples from the few-shot classes is improved.

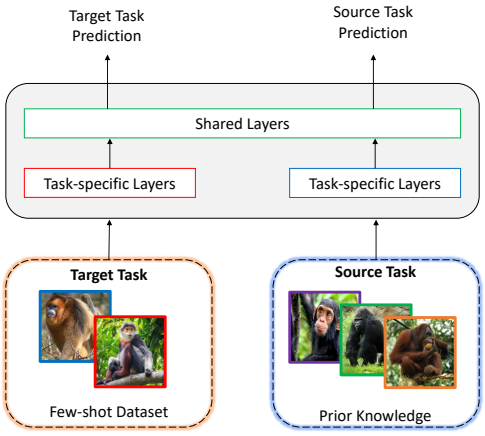

Figure 2.1: Addressing the Few-shot Learning Problem by Multi-task Learning

2.2.2 *Embedding Learning*

Embedding learning, also referred to as distance-metric learning in some of the literature, is an approach that aims at transforming samples $x \in \mathcal{X}$ to a lower-dimensional embedding $z \in \mathcal{Z}$ in which samples that are similar can be easily identified. This approach, ilustrated in Fig. 2.2 is based on three key components: a function $f(.)$ that embeds the training samples, a function $g(.)$ that embeds a test samples, and a similarity function $s(.)$ that measures how similar the embeddings of the test samples and the training samples. The embedding functions $g(.)$ and $f(.)$ are learned from prior knowledge, and in some approaches can learn information from the few-shot dataset. Many approaches have been proposed in the literature such as Matching Networks [29], Prototypical Networks [30], Relation Networks [31], and many others [22].

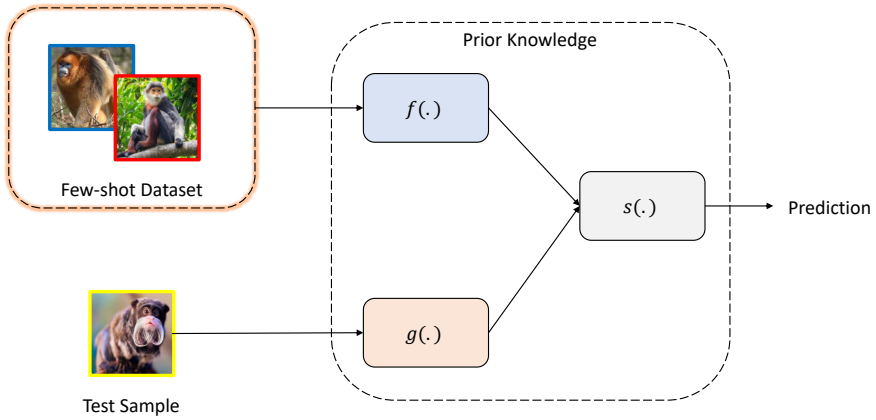

Figure 2.2: Addressing the Few-shot Learning Problem by Embedding Learning

2.2.3 *Refining Existing Parameters*

The aim of this strategy is to find good initialization parameters for the model prior to training on the few-shot dataset. Consider a model with initial parameters θ. Prior knowledge is leveraged from large related datasets that are used to pre-train the model in a supervised manner. This pre-training process results in refined parameters θ_0 that capture general structures from the large source datasets. At this point, the model can now be adapted to the few-shot dataset by training for a few iteration to reach the optimal θ^* that perform well on the few-shot task [32]. This process is illustrated in Fig. 2.3. A similar approach for parameters refined is the aggregation of the weights of several models pre-trained on similar tasks, which can be useful when θ_0 cannot be directly obtained from a very close and large dataset [33]–[35].

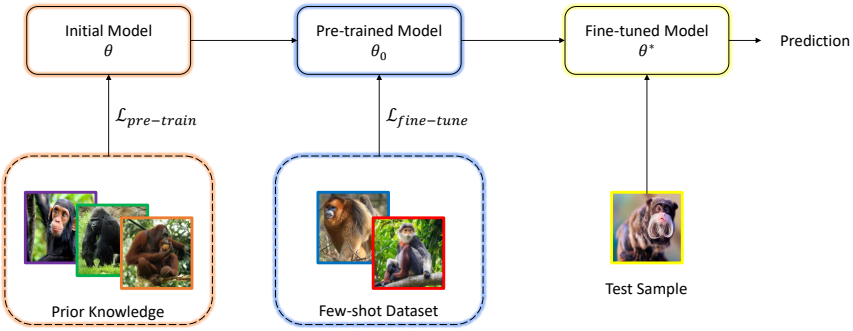

Figure 2.3: Addressing the Few-shot Learning Problem by Refining Existing Parameters

2.3 Few-shot Learning in NLP

2.3.1 *Unsupervised Feature Embedding Representation*

Consider an input text sequence $\mathbf{x}$ composed of tokens $\{x_0, x_1, \ldots, x_L\}$ where L is the sequence length. The aim is to extract from $\mathbf{x}$ a smaller discriminative representation z of size N that carries the necessary information to perform tasks such as classification. As we are dealing with a very limited data regime, x is first transformed into an embedding representation $\mathbf{e} = \{e_0, e_1, \ldots, e_L\}$ where $e_i \in \mathbb{R}^K$ represents the K-dimensional word embedding of the token x_i. The embedding representation $\mathbf{e}$ transfers semantic knowledge of the language, helping deal with the lack of supervised examples in the few-shot dataset. These embeddings can be directly obtained using pre-trained word embedding models such as GloVe [36] or Word2Vec [37]. Finally, a parameter-free operation $\mathcal{P}$ is computed on $\mathbf{e}$ to extract the critical information required for the desired task. This approach, illustrated in Figure 2.4, has been shown beneficial to text classification purposes where limited support examples are available such as Chinese text, [38], biomedical documents [39], and dialectal Arabic text [40].

Several operations for $\mathcal{P}$ have been proposed in the literature. In [41], pooling mechanisms were introduced to obtain z from $\mathbf{e}$. These pooling mechanisms include mean, max, concatenated, and hierarchical pooling.

Mean Pooling: In mean pooling every word contributes to the prediction, hence the pooled representation is obtained through averaging the K-dimensional embeddings:

$$z^{mean} = \frac{1}{L}\sum_{i=0}^{L} e_i \tag{2.1}$$

Max Pooling: Contrary to mean pooling, max pooling filters out unimportant words by extracting salient features from the embedding representation, given the fact that irrelevant words would have smaller amplitudes. Thus, only important keywords

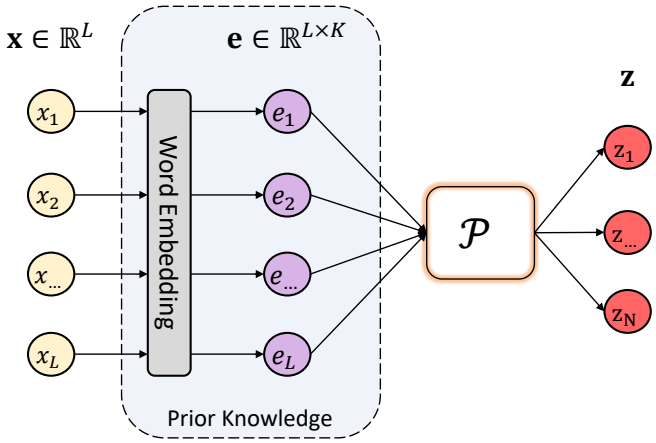

Figure 2.4: Addressing the Few-shot Learning Challenge by Unsupervised Feature Embedding Representation

contribute to the prediction:

$$\begin{aligned} z^{max} &= \max_{i=0}^{L}\{e_i\} \\ z_j^{max} &= \max_{i=0}^{L}\{e_{ji}\} \end{aligned} \tag{2.2}$$

Concatenated Pooling: Mean and max pooling account for different types of information in a text sequence. Concatenated pooling is a combination of both where z^{mean} and z^{max} are concatenated to form the final representation:

$$z^{concat} = [z^{mean}; z^{max}] \tag{2.3}$$

Hierarchical Pooling: Hierarchical pooling is a more sophisticated strategy that takes into account word order and spatial information. In hierarchical pooling, a sliding window of size n is defined and used to compute the mean pooled representation of a set of local windows, over which a global max pooling operation is the computed. The local window is represented by:

$$e_{i:i+1-n} = [e_i, e_i + 1, \ldots, e_i + 1 - n] \tag{2.4}$$

A modified hierarchical pooling mechanism was proposed in [38], where the order of the hierarchical pooling approach was reversed so that local information is first extract using a max pooling sliding window, and then mean pooling is applied for a global representation of the sentence. The modified approach was to better in representing the semantics of the text as well as providing higher performance on documents with large sizes.

2.3.2 *Learning from Label Descriptors*

Consider the task of classifying an input text sequence $\mathbf{x}$ given a pre-defined and fixed set of N labels. We denote by $\mathbf{L} = [L_1, L_2, \ldots, L_N]$ a vector of label descriptions where $L_i = [l_1, l_2, \ldots, l_L]$ is a sequence of tokens that provide a description of the i-th label. When learning form label descriptors, the aim is to leverage prior knowledge of each label's natural language description to better discriminate input samples that belong to labels that rarely or never occur in the dataset. The input $\mathbf{x}$ and label descriptors $\mathbf{L}$ are first transformed into an embedding representation denoted by $\mathbf{e}$ and $\mathbf{u}$ respectively. Each embedding representation is then encoded using a neural layer of choice, resulting in the encoded representations $\mathbf{z}_e$ for the input, and $\mathbf{z}_u$ for the label descriptors. To capture the relevant content in the input text sequence that can provide relevant information for the prediction, label-wise attention is computed between $\mathbf{z}_e$ and $\mathbf{z}_u$:

$$c_i = softmax(\mathbf{z}_e \mathbf{z}_{ui}) \tag{2.5}$$

where c_i measures how informative each part of the input sequence to the i-th label. The prediction is then finally computed at the output layer which takes in takes in $\mathbf{c}$ and $\mathbf{z}_u$ to produce $\hat{y}$. This procedure is illustrated in Figure 2.5.

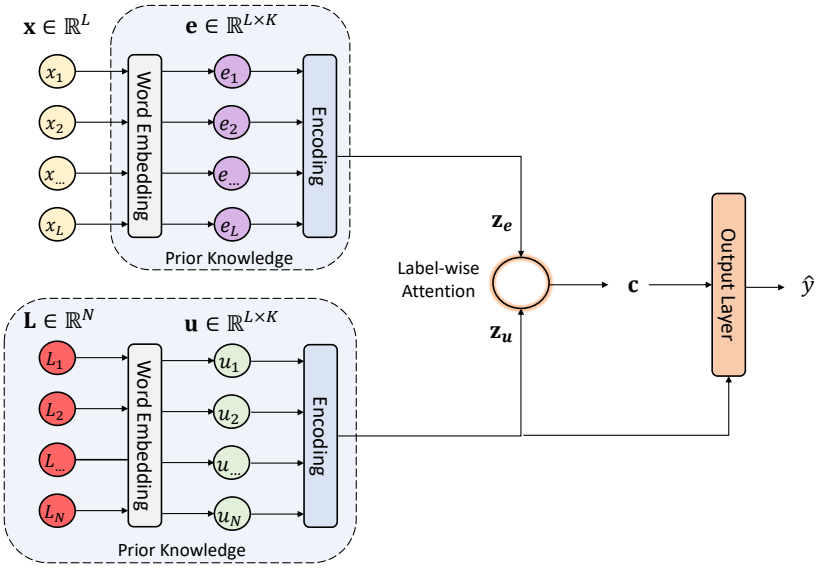

Figure 2.5: Addressing the Few-shot Learning Challenge by Learning from Label Descriptors

This approach has been applied in [42] for few-shot multi-label text classification where graph convolutional neural networks were also used to extract hierarchical information of the medical descriptors that are structured in nature prior to being fed to the output layer. Similar approaches were proposed in [43], [44] for few-shot charge

prediction where the dataset suffered from highly imbalanced and confusing labels, classification of clinical text [45] and legal documents [46], sequence labeling [47].

2.3.3 *Generating with Pre-trained Language Models*

Consider the task of generating a sequence y provided a few m samples of semi-structured data which is in the form of key-value pairs $\{K_i : V_i\}_{i=1}^{m}$ where $V_i = [v_0, v_1, \ldots, v_L]$ is a value sequence of length L that corresponds to the key K_i. The problem has been of particular interest for few-shot natural language generation in task-oriented dialogue agents where the key-value pairs would be extracted slots for domains with limited data samples. Training supervised models to handle these rarely occurring or emerging domains would not yield good performance. Recent attempts in few-shot natural language generation leverage the innate language skill of pre-trained language models to address this challenge [48]. Additionally, pre-trained word embedding representations are used to overcome the small vocabulary size of the few-shot dataset, helping achieve better generalization with tokens unseen during training. In [49], the authors proposed a switch policy framework where the model learns to jointly copy tokens from the key-value pairs and use a pre-trained GPT model as an off-the-shelf generator to form coherent sentences. Authors in [50] propose a table transformation module to model the key-value pairs structure as input to a GPT-2 model. A similar approach for knowledge graph-to-text is proposed in [51] where representation alignment is introduced to bridge the semantic gap between knowledge graph encodings and pre-trained language models.

2.3.4 *Unsupervised Pre-training & Fine-tuning*

This approach aims at finding initialization parameters upon which the model can be fine-tuned with minimal labeled examples to achieve good generalization. Consider the few-shot dataset $D_{few-shot}$ of the target task $\mathcal{T}$ and a corpus D with a large number of unlabeled examples that are much easier to collect than labeled examples for the target task $\mathcal{T}$. We start off by a randomly initialized model with parameters θ_0. This model is then pre-trained in an unsupervised/self-supervised manner using a specific loss function for pre-training $\mathcal{L}_{pre-train}$ to find better initialization parameters θ_p. The model is then adapted to $\mathcal{T}$ by fine-tuning it with a task-specific loss function $\mathcal{L}_{fine-tune}$ to find to best possible task-specific parameters θ^*. The procedure is illustrated in Figure 2.6.

This approach has been applied in the literature for various applications. In [52], the authors address low-resource response generation by unsupervised pre-training on large corpora of Chinese text. The pre-training step helps the model learn semantic and syntactic information about the language before being fine-tuned on the response generation task using a smaller set of utterance-response pairs. Similar approaches have been applied in [53] for task-oriented dialogue generation where an additional step of pre-training on labeled examples from domains with abundant data is applied before

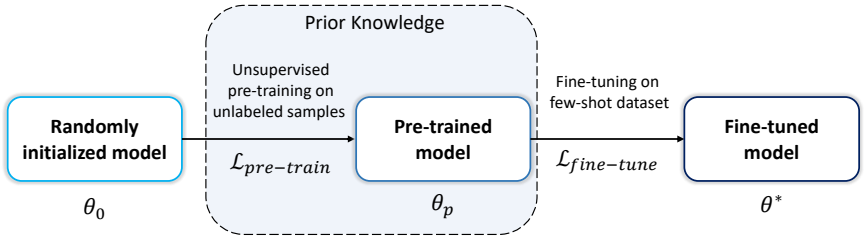

Figure 2.6: Addressing the Few-shot Learning Challenge by Unsupervised Pre-training & Fine-tuning

fine-tuning on the few-shot emerging domains where limited samples are available. Similar approaches that benefit from pre-trained model representations for text generation in task-oriented dialogue agents are presented in [54], [55] Few-shot intent classification is done by leveraging conversational knowledge from a model pre-trained on large amounts of conversational in [56], or from pre-trained language models in [57].

2.3.5 *Meta-Learning*

Meta-learning is commonly referred to as "learning to learn" where the aim is to improve a learning algorithm over several learning episodes. In traditional learning, a learning algorithm is trained to solve a task using a task-specific dataset and an objective function. In meta-learning, a meta-learning extracts meta-knowledge from multiple source tasks to update an inner learning algorithm in a way to learn a model can achieve a certain objective.

In conventional supervised learning, given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{m}$ we are interested in finding a hypothesis function h_θ parameterized by θ that minimizes a loss function $\mathcal{L}$ that measures the error between the ground-truth labels y and the predictions produced by h_θ. This function is learned by solving:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\mathcal{D}, \theta, w) \tag{2.6}$$

where w specifies the assumptions on how the model will learn, which includes the optimization strategy, the initialization parameters, the function class for h, or the entire learning algorithm. In conventional learning, models are trained from scratch on D with pre-defined knowledge of w. The main issue with the pre-definition of w is that it could affect the model's performance and sample efficiency. In few-shot learning settings, using meta-learning approaches to extract meta-knowledge of w from source tasks can help improve performance on the target few-shot task. In this regard, we define a meta-training as the stage of learning w from a set of K source datasets D_{source} as follows:

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{K} \mathcal{L}(D_{source}^{(i)}; w) \tag{2.7}$$

In the meta-testing stage, the base model is trained on the few-shot dataset D_{few_shot} given prior knowledge of $w*$ extracted by the meta-learner. Hence, the meta-testing stage becomes:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L}(D_{few-shot}; \theta, w^*) \tag{2.8}$$

Therefore, by leveraging the knowledge gained of w in the meta-training stage, performance on the few-shot dataset can be improved compared with from-scratch training. In this regard, meta-learning approaches have been adopted in various NLP applications including neural machine translation [58], dialogue-state tracking [59], user intent classification [60], dialog generation [61], event detection [62], named entity recognition [63], and others [64], [65].

CHAPTER 3

MESSAGE-RESPONSE DATASET FOR MSA AND BASELINE SEQ2SEQ MODEL

3.1 Objectives

Despite the various works presented in the literature on open-domain empathetic chatbots for English, no work has previously addressed the problem of building such models for the Arabic language. An important reason is the scarcity of resources available for Arabic compared with the English language, including datasets and pre-trained language generation models. In this work, we create a dataset of empathetic utterance-response pairs in MSA by translating the EmpatheticDialogues dataset available for English. We train a Seq2Seq model with Long Short-Term Memory (LSTM) units on the translated dataset. The developed model successfully exhibited empathetic behavior and provided emotional responses to the input of users in MSA.

3.2 Related Work

3.2.1 *Open-Domain English Chatbots*

Conventional Approaches: Early attempts in building open-domain empathetic bots relied on developing NLP modules and a dialog manager to switch between modules [66]–[68]. The general architecture of the conventional approach is illustrated in Fig. 3.1. First, a user understanding module handles in the user input utterance, where multiple classification models are used to identify the user's intent and emotional state. This module also identifies the current context of the chatting session and keeps track of user-related information to create a user profile. The information collected from the user understanding module is then passed on to a response generation module. At this stage, the response is obtained either through retrieval-based or generative approaches. To maintain a consistent personality for the chatbot, a personality setting module is used to control the language of the chatbot. Additionally, to make sure an bias and inappropriate responses are provided by the bot, an ethical design module is used to

handle such issues. However, such approaches remained limited in their capabilities and fail to generalize beyond a specific set of domains. In this regard, such architectures have been extremely useful in industrial applications, such as task-oriented chatbots like Apple's Siri or Amazon's Alexa, where they are expected to perform very well in specific domains only.

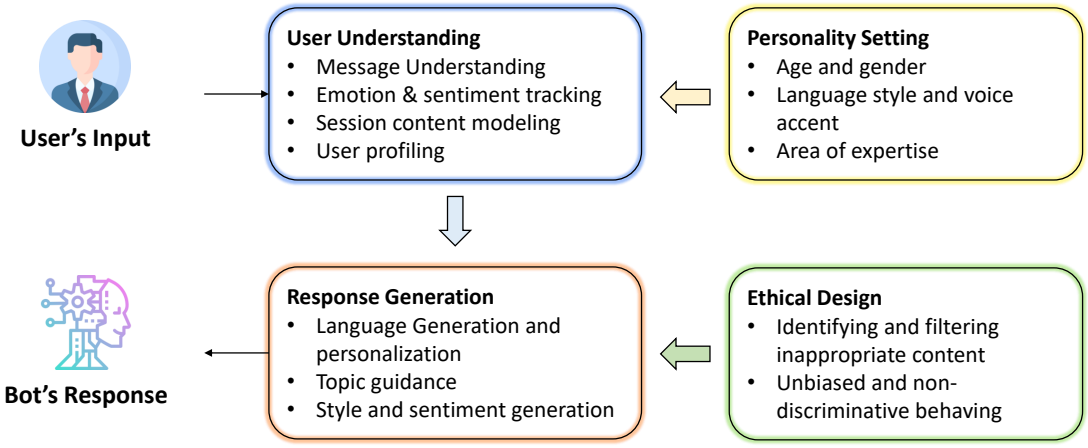

Figure 3.1: Architecture of the conventional approach for open-domain empathetic bots based on NLP modules

End-to-End Generative Approach: With the advances in AI techniques and computation power, researchers recently showed that open-domain conversational models developed using end-to-end neural network-based approaches, such as Sequence-to-Sequence (Seq2Seq) models, generalize well to unforeseen domains without relying on complex architectures of hand-crafted modules [6]. These approaches, however, require training on large corpora of open-domain conversational data [69]. Additionally, Seq2Seq models benefit from the availability of pre-trained language generation models that can be fine-tuned for downstream generation tasks [70]–[72]. As such, those models can acquire a solid representation of the language, and of language generation skills before being adapted to the task-specific task such as dialogue response generation. The availability of massively pre-trained language generation models also help address the issue of data scarcity whereby less task-specific labeled data would be needed to reach reasonable performance.

Empathetic Open-Domain Conversational Models: English empathetic chatbots have been of interest over the last few years. Recently, the first dataset for empathetic conversations dubbed EmpatheticDialogues was introduced by *Rashking et al.* [9]. In their work, the authors gathered the dataset through the use of Amazon Turk Workers and then implemented retrieval-based and generative-based models. Overall, they observed higher levels of empathy in the chatbot's responses compared with models trained on conventional non-empathetic datasets. The same dataset would later be used as the main benchmark for assessing empathetic models. For instance, in [11], the authors proposed an improved model which employed a Generative Pre-trained

Transformer (GPT). This model was pre-trained on the BooksCorpus dataset [73] that contains over 7000 unpublished books, thus improving the Natural Language Understanding (NLU) ability of the transformer. They also pre-trained on the PersonaChat dataset [74] to give the chatbot a certain persona and enhance its engagingness. Following this pre-training procedure, the model was fine-tuned on EmpatheticDialogues with results showing significant improvements in the empathetic responding capability of the model.

In a different approach [10], the authors modeled empathetic responding as a reinforcement learning problem where they defined a reward function for a Seq2Seq model based on Gated Recurrent Units (GRU) and attention. Their approach named "Sentiment Look-ahead" is also shown to be effective in generating empathetic responses when tested on the EmpatheticDialogues dataset. In [75], the authors approached the problem from a different perspective, splitting it into an emotion recognition and a response generation problem. Inspired by Affect Control Theory, they map every user sentence to an EPA (Evaluation Potency Activity) vector using a BiLSTM network with attention and then prescribe a corresponding EPA response vector which they use for conditioning the response generation. Both Conditional Variable Auto Encoders and Seq2Seq models are considered for the generation. They are seen to yield similar results. In [76], the authors also make use of a Seq2Seq model for their chatbot's general chitchat. They represent empathy through the use of two empathy vectors, one which represents the user (including sentiment, opinion, and contextual information) and one which represents the chatbot (including its opinion and personality). They condition the decoder on these empathy vectors and learn the best replies for each situation using data from interactions with over 660 million users.

3.2.2 *Arabic Chatbots*

Arabic is a complex language and thus the development of Arabic chatbots has been a great challenge to the research community. To date, only a handful of works have attempted to build Arabic chatbots. One such work is ArabChat: a rule-based chatbot capable of pattern matching and providing suitable answers to queries by the users [20]. Another work is BOTTA, a retrieval-based model supporting specifically the Egyptian dialect [19]. For the medical domain, Ollobot is another rule-based chatbot which presents health tracking and support [21]. Overall, in[77], it was seen that Arabic chatbots are still in their infancy. Their development being mainly hindered by a lack of available datasets. Some works have managed to break through this limitation by leveraging translation tools: an example is the question answering system [78], another one is the Arabic language model [18]. The success of these works, as well as the work in [79] demonstrate the potential of neural models in understanding the Arabic language and motivates us to look into neural solutions for the open challenge of Arabic empathetic response generation.

3.3 Proposed Method

3.3.1 *Arabic Dataset for Empathetic Chatbots*

The proposed model requires training on a dataset of empathetic conversations. A sample input in this dataset would be a statement of a speaker describing personal experience in which they felt a specific emotion. The corresponding output would be the empathetic response of a listener, which infers the emotional state of the speaker and provides an appropriate reply. The proposed model needs to be trained on these input-output pairs so that it could generate human-like empathetic responses.

Since no such dataset is available in the Arabic language, we translated the EmpatheticDialogues dataset [9], which is the only available dataset in English for building empathetic chatbots. EmpatheticDialogues consists of 24,850 English conversations obtained via crowd-sourcing. These conversations are between a speaker that describes a certain situation they went through and a listener who infers the emotional state of the speaker and provides a suitable emotional response, thus creating an empathetic dialogue. We make use of the Googletrans[1] API to perform the translations from English to Arabic. Data utterance-response pairs for various emotional contexts are provided in Table 3.1.

Emotion	**Excited**
Utterance	في الأسبوع الماضي كنا نشاهد كأس العالم ، وها هي كرواتيا تغلبت على البلد المضيف، روسيا
Response	رأيت تلك اللعبة. لقد فازوا بركلات الترجيح!
Emotion	**Furious**
Utterance	تخيل، لقد طلبت البطاطس المقلية و لكن تم تقديم البرغر بدلا من ذلك!
Response	إنها خدمة سيئة للغاية. ماذا فعلت؟ هل اشتكيت للمدير؟
Emotion	**Embarrassed**
Utterance	خرجت في نهاية الأسبوع الماضي و تعرضت لحادث كبير. خمن ماذا جرى
Response	هل أنت بخير؟ عليك أن تخبرني بما حدث

Table 3.1: Samples of utterances and empathetic responses from the ArabicEmpatheticDialogues dataset for three emotion labels: Excited, Furious, and Embarrassed

To evaluate the quality of the dataset, we chose 100 random translated samples and compared them with the original English samples to assess the quality of the translation. Our interest in the dataset is not to obtain accurate translations, but rather to create dialogues that are meaningful even if they were not perfect translations. As a result, our evaluation of the dataset focused on checking whether the translated conversation makes sense in Arabic. The results indicated that only 6 of the 100 randomly chosen samples were found to be unreasonable while the rest of the samples were deemed reasonable. Therefore, we considered the dataset to be of high quality for the purpose

[1] https://pypi.org/project/googletrans/

of training the proposed empathetic conversational model. A few unreasonable samples are shown in Table 3.2. Such poor translations are mainly due to idioms of the English language, where the individual words do not represent the literal meaning. For instance, by looking at the sample "Planning out my new home has turned out to be a blast!" the word "blast", in the context of the sentence, means "exciting" while its literal meaning is "explosion". Another reason for unreasonable translations are slang words, which are commonly found in informal conversations. These types of errors are rare in the generated conversation dataset and the translation system was thus deemed to be sufficiently accurate (94%) for the purpose of model development.

Planning out my new home has turned out to be a blast!
تبيّن أن التخطيط لمنزلي الجديد كان إنفجارا !
I suppose you do have a point there
أعتقد أن لديك نقطة هناك

Table 3.2: Examples of unreasonable translations.

3.3.2 *Proposed Arabic Encoder-Decoder Model*

The purpose of the model is to infer an emotional state in an input sequence, that is the user's statement, and generate a sequence in Arabic representing the empathetic response that the chatbot needs to reply with. The proposed model, illustrated in Fig. 3.2, is a Seq2Seq model with LSTM units combined with Attention. The components and parameters of the proposed model were obtained following a process of hyperparameter tuning that determined the combination of choices that will deliver the best performance on the validation set. The hyperparameters tuned were the number of encoder/decoder layers (1, 2, or 3 layers), unit type (LSTM or GRU), embedding dimensions (100, 200, or 300), and choice of optimization algorithm (Stochastic Gradient Descent (SGD), Adam, or Adagrad). After trying all combinations and comparing performance on the validation set, the resulting choices of hyperparameters were two layers for each the encoder and decoder, LSTM units, an embedding dimension of 500, and SGD as the optimizer during training and validation. A dropout probability of 0.3 was chosen after each layer to avoid over-fitting.

We consider the empathetic conversations to be alternating sequences between the user and the chatbot. Let $w = [w_1, w_2, \ldots, w_{n_x}]$ be the input one-hot representations of a sequence of n_x words, corresponding to the utterance said by the user. We use the Farasa [80] Arabic text processing toolkit for pre-processing and tokenizing Arabic sentences. The obtained tokens are then fed into an embedding matrix $E \in \mathbb{R}^{d \times V}$ where d is the dimension of the embedding vector, and V is the vocabulary size. We set d to be 500 and obtain a vocabulary size V of 12900. The output of the embedding layer results in $x = [x_1, x_2, \ldots, x_{n_x}]$ where x_i is the embedding vector of the i-th word

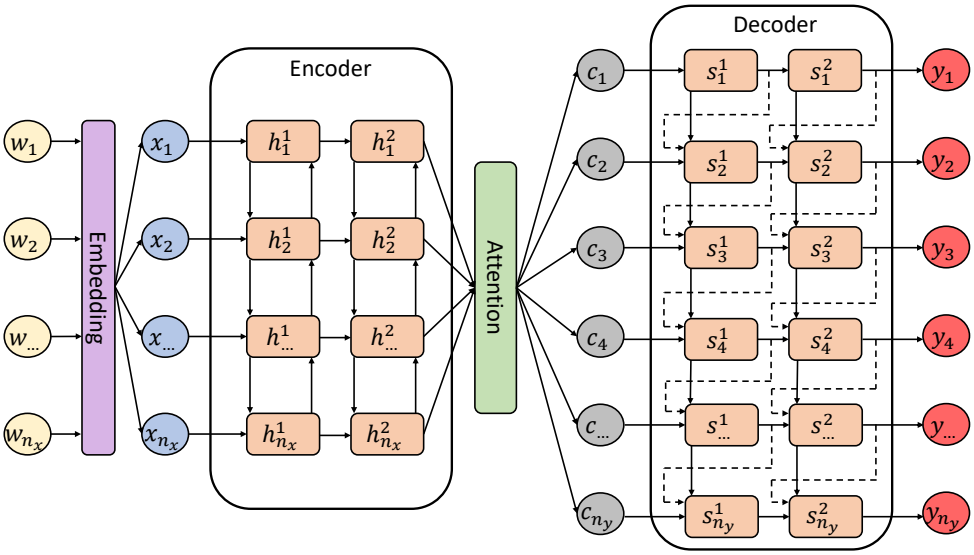

Figure 3.2: Architecture of the proposed Seq2Seq model with Attention

w_i. The target output sequence is the sentence containing an empathetic response by the chatbot and which we represent by $y = [y_1, y_2, \ldots, y_{n_y}]$.

The encoder consists of two bidirectional layers with LSTM units for better extraction of complicated features. Each unit computes a hidden state h_i^l where l is the layer index. To avoid the problem of fixed-length vectors in encoder-decoder models [81], we used an attention mechanism which generates a context vector $c = [c_1, c_2, \ldots, c_{n_y}]$ given the hidden states h_i^2 from the second layer of the encoder. The context vector c is then fed as input to the decoder that consists of two layers with LSTM units. Each unit computes the hidden state s_i^l. The sequence y representing the empathetic response can then be generated by the second layer of the decoder, where a decoding strategy is used such as beam search or random sampling. The mathematical details of the model are provided here for completeness.

Encoder: The encoder is formed by two stacked layers of bidirectional LSTM units (BiLSTM) that compute the encoder hidden states denoted by h_i^l where l is the layer index. The first layer reads the input embeddings x and computes the hidden state $h_i^1 = [\overrightarrow{h^1}_i; \overleftarrow{h^1}_i]$ in both directions as follows:

$$\begin{aligned} \overrightarrow{h^1}_i &= \overrightarrow{\text{LSTM}}(\overrightarrow{h^1}_{i-1}, x_i) \\ \overleftarrow{h^1}_i &= \overleftarrow{\text{LSTM}}(\overleftarrow{h^1}_{i+1}, x_i) \end{aligned} \tag{3.1}$$

The obtained hidden states of the first layer are then fed as input to the second layer to compute $h_i^2 = [\overrightarrow{h^2}_i; \overleftarrow{h^2}_i]$ as follows:

$$\begin{aligned}\overrightarrow{h^2}_i &= \overrightarrow{\text{LSTM}}(\overrightarrow{h^2}_{i-1}, \overrightarrow{h^1}_{i-1}) \\ \overleftarrow{h^2}_i &= \overleftarrow{\text{LSTM}}(\overleftarrow{h^2}_{i+1}, \overleftarrow{h^1}_{i+1})\end{aligned} \tag{3.2}$$

Attention: We employ the attention mechanism where attention weights α_{ji} are assigned to each hidden state h_i^2 obtained by the encoder. These weights are computed by:

$$\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_k^{n_x} \exp(e_{ki})} \tag{3.3}$$

where the energy e_{ji} associated with each weight α_{ji} determines how significant an encoder state h_i is to a decoder state s_{j-1} in generating the next state s_j. The energy is computed using the alignment model given by:

$$e_{ji} = f_{NN}(s_{j-1}, h_i) \tag{3.4}$$

where f_{NN} denotes a regular feed-forward neural network that is trained simultaneously with the rest of the system. The context vector c_j can now be computed by the weighted sum of α_{ji} and h_i for $j = 1, \ldots, n_x$ as follows:

$$c_j = \sum_{j=1}^{n_x} \alpha_{ji} h_i \tag{3.5}$$

Decoder: The decoder consists of two stacked layers of LSTM units that compute the decoder states s_j^l as follows:

$$\begin{aligned}s_j^1 &= \text{LSTM}(c_j, s_{j-1}^1, s_{j-1}^2) \\ s_j^2 &= \text{LSTM}(s_j^1, s_{j-1}^2, y_{j-1})\end{aligned} \tag{3.6}$$

Hence, the next word in the generated empathetic response y_j can be predicted given the previously predicted words $y_1, y_2, \ldots, y_{j-1}$ and the context vector c.

$$p(\mathbf{y}) = \prod_{j=1}^{n_y} p(y_j / y_1, y_2, \ldots, y_{j-1}, c) \tag{3.7}$$

where $\mathbf{y} = y_1, y_2, \ldots, y_{n_y}$.

3.4 Experiments & Results

3.4.1 *Experimental Setup*

The created dataset contains around 35K samples which are split into 80% for training, 10% for validation, and 10% for testing. We train the proposed Seq2Seq model for

three different embedding dimensions (d) of 100, 300, and 500, to explore how this dimension will influence the performance of the model given the vocabulary size we have. SGD was chosen as the optimization algorithm during training. Additionally, we applied a dropout probability of 0.3 after each layer. The models were developed using the OpenNMT [82] toolkit which is commonly used for neural sequence learning.

3.4.2 *Model Training and Evaluation*

During the training and validation process, the models are evaluated using the Perplexity (PPL) automated metric. The curves in Fig. 3.3 show the variation of the validation PPL over 8000 training steps, for the three choices of d. As observed in Fig. 3.3, the model with $d = 500$ achieved the best value for the PPL on the validation set, reaching nearly 30, while the models with $d = 100$ and $d = 300$ showed a validation PPL around 50. The summary performance of these models is reported on the test set in Table 3.3, where beam search is used at inference time. We use the BLEU score as an additional metric for evaluation. The model with an $d = 500$ outperforms the rest of the models by achieving the highest BLEU score of 0.5 and lowest PPL of 38.6 on the test set. Given these obtained values for the PPL and the BLEU score, the model delivered state of the art performances for Arabic. The state of the art models for English reached a PPL level close to 10. However, the achieved results for Arabic are considered as very good given the relatively small size of the dataset used and the more complex nature of the Arabic language. Reaching even better PPL and BLEU score levels would require more data samples to learn from. A possible solution could be pre-training on larger conversational datasets in Arabic, that would contain hundreds of thousands of samples, and fine-tuning on the empathetic conversations dataset.

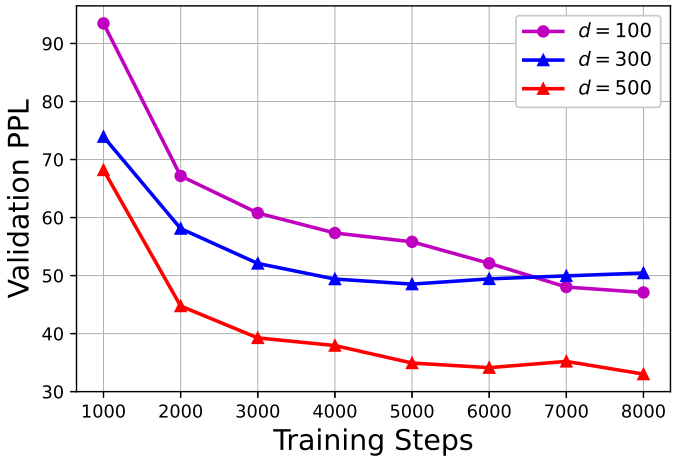

Figure 3.3: Validation PPL curves for several word embedding dimensions d

Embedding Dimension d	PPL	BLEU
100	53.5	0.11
300	48.7	0.32
500	**38.6**	**0.50**

Table 3.3: Performance of the models on the test set in terms of PPL and BLEU score.

3.4.3 *Evaluation by Human Annotators*

Automated metrics such as PPL and BLEU score do not capture all aspects of performance of models in sequence generation and cannot be used alone to judge the quality of the responses generated since they don't always correlate with human judgment. This problem is especially applicable to empathetic chatbots, where no metric exists to evaluate how empathetic the response generated is. Thus, human ratings are an important part of the overall evaluation. To this end, we collected ratings from 50 speakers of the Arabic language. The raters were given samples from each model and were asked to rate them in terms of Empathy, Relevance, and Fluency, by answering the following questions:

- Empathy: Did the response show an ability of inferring the emotions in the given context?
- Relevance: How relevant was the generated response to the given input context?
- Fluency: How understandable was the generated response from a language perspective?

The raters were asked to rate responses on a scale from 0 to 5, where 0 conveys terrible performance and 5 conveys excellent performance. The average of the obtained human ratings are reported in Table 4.3 for each model. We experimented with two decoding strategies at inference time, which are namely beam search and random sampling. The model with $d = 500$ and which uses beam search in sequence generation achieved the highest ratings in all of the specified metrics. These ratings suggest that the model exhibits state of the art performance for Arabic with average levels of Empathy and Fluency reaching of 3.7 and 3.92 respectively. However, the Relevance metric was at 3.16 reflecting that the model did not always stay on topic while generating responses. Hence, it can be deduced that the model provides fluent and empathetic responses, but could go off topic in some cases and respond with irrelevant statements.

For further analysis of the model's performance, we show in Table 3.5 examples of generated responses by the model on a set of context sentences that were not included in the training dataset. We also compare these generated responses for the same model using the random sampling decoding strategy. Several points can be deduced by analyzing the generated responses in Table 3.5. We notice that even though beam search provides fluent and empathetic responses, the responses from beam search are limited to a few choices. For instance, we observed that the tokens (يا للروعة) were repeated several times for different contexts. Sometimes, a full sequence is repeated such as

Decoding Strategy	Embedding Dimension d	Empathy	Relevance	Fluency
Beam Search	100	2.24	1.96	3.08
	300	2.5	2.26	3.03
	500	**3.70**	**3.16**	**3.92**
Random Sampling	100	2.04	1.68	2.44
	300	2.03	1.69	2.57
	500	2.40	1.92	2.80

Table 3.4: Average of human ratings collected for several embedding dimensions and decoding strategies.

(يا للروعة، يجب أن تكون فخور جدا). This repetitive behavior makes the model seem limited by only a few sequences to choose from and gives the impression that it is not capable of generating more general sequences. In few cases, the response did not make perfect sense or the response went totally off-topic. This issue is commonly encountered when using beam search even for English models. Another drawback of beam search is the heavy computational load it imposes since it needs to perform exhaustive search.

With random sampling, the next token in the sequence is generated based on the probability distribution obtained by the softmax function. This approach, as seen in Table 3.5, generated lengthy sequences and avoided being too generic as in beam search, and thus offered more richness in the response choices. However, the human ratings for the random sampling approach dropped significantly compared with the models using beam search. This is because in random sampling, many unlikely tokens had an increased probability of being generated, and much more training data would be needed to learn from for the performance to improve.

Additionally, it is noticed from Table 3.5 that when the context is a simple question that infers no emotions in the speaker such as (ماذا تفعل؟) or (مرحبا كيف حالك؟), the model still provided a response with an unnecessary emotional state. This incapability of the model to generate regular chit-chat responses is observed when using either of the decoding strategies, and is mainly due to it being trained merely on a dataset of empathetic conversations. Hence, it will always opt to generate an empathetic response to any context it receives. Pre-training the model on standard Arabic conversational datasets, and then fine-tuning on our proposed Arabic empathetic dialogues dataset should help alleviate this problem.

Context	Generated Response	
	Beam Search	Random Sampling
مرحبا كيف حالك ؟	كنت سعيدا جدا في ذلك اليوم	هذا أمر محزن هل أنت متوتر للذهاب إلى مكان ما ؟
هل زرت لاس فيجاس من قبل ؟	نعم، لقد كان الوقت متأخرا جدا	هذا أمر محزن للغاية آمل أن يكون لديك الكثير من المرح
ماذا تفعل ؟	كنت خائفة من السير إلى المنزل	و لكن علي أن يتعافى مرة أخرى
اليوم يوم سيئ جدا	ماذا حدث ؟	لا يزال أتذكرني فقط من قبل
أنا سعيد جدا	هذا مثير للإعجاب، ماذا فعلت ؟	هذا لطيف جدا لهم ما الذي جعلك تعمل بسرعة ؟
لقد ربحت مباراة كرة القدم اليوم	يا للروعة، يجب أن تكون فخورا جدا	هل فكرت في نفسها ؟
خلال عيد الميلاد قبل بضع سنوات لم أحصل على أي هدايا	يا للروعة، هذا يبدو مثير للإعجاب	هذا فظيع، نفس الشعور
لقد تلقيت ترقية في عملي اليوم	يا للروعة، يجب أن تكون فخورا جدا	هذا يبدو رائع يجب أن تكون فخورا حقا بنفسك

Table 3.5: Sample generated responses by the proposed model. Generated responses are shown using both beam search and random sampling decoding strategies at inference.

CHAPTER 4

BERT2BERT CONVERSATIONAL MODEL: LEARNING ARABIC LANGUAGE GENERATION WITH LITTLE DATA

4.1 Objectives

To address the challenges of small dataset size and lack of conversational resources, in terms of datasets and pre-trained models, we propose a transformer-based encoder-decoder model initialized with AraBERT [18] pre-trained weights. This work extends the English BERT2BERT architecture [83] to Arabic response generation. We fine-tune our proposed model on the limited-sized dataset of empathetic responses in Arabic. By using the pre-trained weights of the AraBERT language model to initialize the encoder and decoder, our proposed BERT2BERT model leverages knowledge transfer and shows enhanced performance in empathetic response generation compared to the baseline Bi-LSTM model proposed in Section **??** both in terms of Numerical Evaluation and Human Ratings.

4.2 Related Work

Recently, the first empathy-driven Arabic conversational model was proposed in [84] that released ArabicEmpatheticDialogues, a dataset of Arabic utterances and their corresponding empathetic responses. The authors trained a Seq2Seq model with bidirectional LSTM units on the dataset. While the model succeeded in generating empathetic responses, it showed an average Relevance score which indicates that the responses can sometimes go off-topic and may not be suitable responses for the emotional context of the input utterance. The limitations of this work were mainly due to the limited size of the dataset. In this work, we adopt the BERT2BERT architecture [83] and leverage the

pre-trained AraBERT [18] model to improve the performance of empathetic Arabic conversational models.

4.3 Proposed Method

4.3.1 *Proposed BERT2BERT Model*

Our proposed model for Arabic empathetic response generation is a transformer-based Seq2Seq model [85], which has been shown to boost performance on a several Seq2Seq tasks [86], [87]. However, such an architecture would require massive pre-training before being fine-tuned on the desired task [88]. It was shown by [83] that warm-starting the transformer-based encoder-decoder model with the checkpoints of a pre-trained encoder (e.g. BERT) allows the model to deliver competitive results in sequence generation tasks while skipping the costly pre-training. Inspired by this idea, and due to the unavailability of Arabic conversational datasets that can be used for pre-training, we adopt the BERT2BERT architecture [83], and warm-start the encoder and decoder with the AraBERT checkpoint [18]. The encoder-decoder attention is randomly initialized. The architecture of the proposed model is illustrated in Figure 4.1.

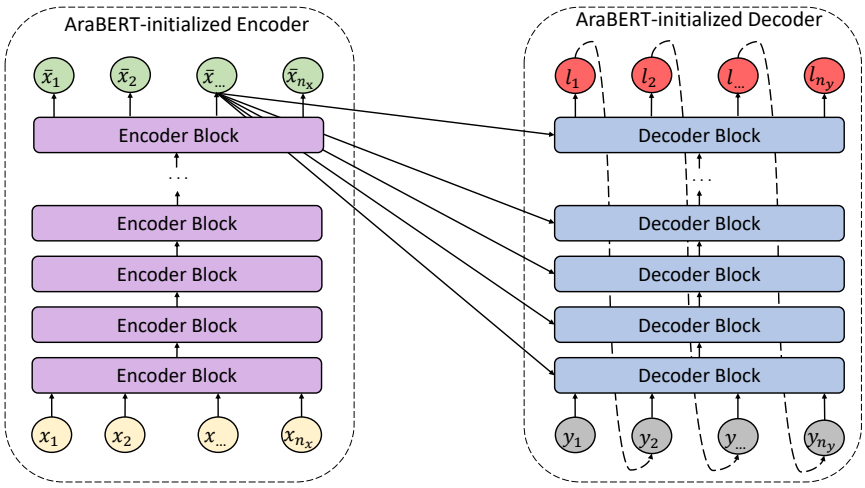

Figure 4.1: Architecture of the proposed BERT2BERT model initialized with AraBERT checkpoints for Arabic empathetic response generation.

The input to the proposed model is a sequence $x = [x_1, x_2, \ldots, x_{n_x}]$ of one-hot representations with a length of n_x tokens, chosen to be 150. This sequence is fed as input to an AraBERT initialized encoder. At the decoder side, the model generates an empathetic response represented by a sequence $y = [y_1, y_2, \ldots, y_{n_y}]$, where the maximum output length n_y is also specified to be 150. We optimize the log-likelihood loss over the output tokens.

To generate empathetic responses from our model, we adopt the Top-K Sampling scheme [89] where, at each time step, the model randomly samples the K most likely

candidates from the probability distribution of all words in the vocabulary. This decoding strategy has been found more effective than conventional approaches such as beam search, which tends to yield common responses found repetitively in the training set or similar, slightly-varying versions of the same high-likelihood sequences [90].

4.3.2 *Dataset*

We use the ArabicEmpatheticDialogues dataset [84] which was translated from the English version introduced in [9]. ArabicEmpathicDialogues contains 36,628 samples of speaker utterances and their corresponding empathetic responses in Arabic. Each sample is also labeled with the emotion of the speaker's utterance. Three examples from the dataset for three different emotion labels are provided in Table 3.1. By training a sequence generation model on the samples of utterances and their corresponding responses from the dataset, the model will be able to infer the emotions in input utterances and provide suitable empathetic responses. Thus, the empathetic capability of the model would be enhanced.

The dataset is originally labeled with 32 emotion labels, many of which are very similar such as "joyful" and "content", or "angry" and "furious". To reduce the number of classes, we follow the tree-structured list of emotions defined in [91] to map the 32 emotion labels to their 6 primary emotions which are "Joy", "Surprise", "Love", "Surprise", "Anger", and "Fear".

To reduce lexical sparsity, the utterances and responses in the dataset are segmented using the Farasa segmenter [80]. Given the morphological complexity of the Arabic language, segmentation is an important pre-processing step that can greatly enhance the performance of neural-based sequence generation models. An example of this process is shown in Table 4.1. By performing segmentation, the vocabulary size is drastically reduced from 47K tokens to around 13K tokens.

Pre-Segmentation
أنا فخور جدا بإبنتي لقد تخرجت للتو من كلية الهندسة
Post-Segmentation
أنا فخور جد+ا ب+ إبن +ت +ي لقد تخرج +ت
ل+ ال+ تو من كلي +ة ال+ هندس +ة

Table 4.1: Example of an Arabic utterance segmentation using Farasa.

4.4 Experiments & Results

We evaluate the proposed BERT2BERT model in comparison to three benchmark models. We conduct numerical as well as human evaluation of the different conversational models.

4.4.1 *Benchmark Models*

We train several neural-based sequence generation models on the ArabicEmpathetic-Dialogues dataset and consider them as benchmarks for performance comparison. The benchmark models are denoted as follows:

Baseline: The baseline model, illustrated in Figure 4.2, is a Seq2Seq Bi-LSTM model with Attention following the prior state-of-the-art model proposed in [84].

EmoPrepend: In this setup, illustrated in Figure 4.2, we prepend the emotion label to each utterance before feeding it as input to the baseline model described above, and we denote this approach as EmoPrepend. This allows us to add supervised information to the data, without having to introduce any modifications to the architecture. The existing emotion labels have been prepended to the utterances in the train and validation sets. For the test set and at inference, we fine-tune AraBERT for emotion classification using the utterances and their labels in the dataset. The fine-tuned AraBERT model is then used as an external predictor to classify the emotion in the utterance and prepend it as a token before being used as an input to the EmoPrepend model. We note that the step of grouping emotion labels into 6 main labels, as discussed in Section 3, makes the emotion classification task easier.

BERT2BERT-UN: which stands for BERT2BERT-Uninitialized. This model is a regular transformer-based encoder-decoder model that shares the same architecture of the BERT2BERT model shown in Figure 4.1, but **is not** initialized with AraBERT pre-trained weights.

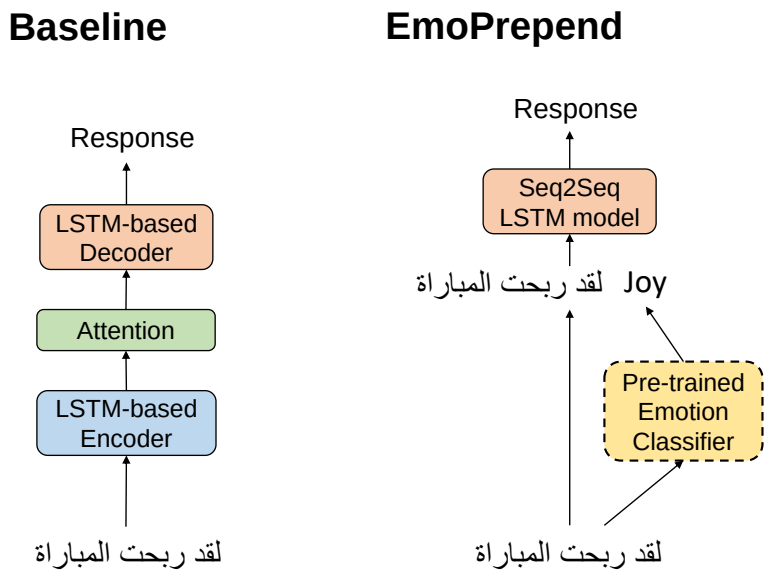

Figure 4.2: Architectures of the Baseline and EmoPrepend models used for comparative evaluation against the proposed BERT2BERT model.

4.4.2 *Experimental Setup*

The proposed BERT2BERT model was developed using the Huggingface transformers library[1]. We train the model for 5 epochs with a batch size of 32[2]. Model training was done on a 16GB V100 NVidia GPU. The Baseline Bi-LSTM Seq2Seq [84], EmoPrepend, and BERT2BERT-UN benchmark models were developed using the OpenNMT Library [82]. A online demo is also available[3].

Dataset Partitioning: All models were trained and evaluated on common data splits of the ArabicEmpatheticDialogues. We randomly partitioned the dataset into 90% training, 5% validation, and 5% testing using a seed of 42.

4.4.3 *Numerical Evaluation*

Table 4.2 summarizes the perplexity (PPL) and Bilingual Evaluation Understudy (BLEU) scores for the proposed and benchmark models when evaluated on the test set. It is clear from the numerical evaluation results that the proposed BERT2BERT model consistently outperforms the benchmark models. This is reflected through both a lower PPL score and a higher BLEU score.

Model	PPL	BLEU
Baseline [84]	38.6	0.5
EmoPrepend	24.1	3.16
BERT2BERT-UN	159.8	0.1
BERT2BERT	**17.0**	**5.58**

Table 4.2: Performance of the models on the test set in terms of PPL and BLEU score.

With EmoPrepend, the addition of supervised information in the form of prepended emotion labels showed performance improvements in comparison to the Baseline model, reflected by an increase in 2.6 BLEU points and a reduction of 14.5 points in the PPL score. Nevertheless, the PPL score of EmoPrepend at 24.1 is still considered high and could potentially lead to sub-optimal performance. BERT2BERT showed significant performance improvements in comparison to the baseline Seq2Seq Bi-LSTM, highlighted by a much reduced PPL value of 17.0 and an increase in 5 BLEU points. BERT2BERT also achieved better scores than the EmoPrepend model.

The BERT2BERT-UN model resulted in a very high PPL score of 158.9 and very low BLEU score of 0.1. These poor results are due to the nature of transformer networks that require huge amounts of data samples to deliver good performance. The initialization of the BERT2BERT with pre-trained AraBERT weights showed very significant enhancements compared with the uninitialized BERT2BERT-UN model. This

[1]https://github.com/huggingface/transformers

[2]https://github.com/aub-mind/Arabic-Empathetic-Chatbot

[3]https://huggingface.co/spaces/tareknaous/arabic-empathetic-response-generation

performance boost provided by the BERT2BERT model is expected given the fact that AraBERT's initialization parameters have been pre-trained on a massive 24 GB Arabic corpus.

The numerical results achieved by the BERT2BERT model are particularly impressive since, despite the limited size of the ArabicEmpatheticDialogues dataset, BERT2BERT was able to leverage knowledge transfer through fine-tuning to achieve state-of-art performance on the task of open-domain empathetic response generation in Arabic without requiring additional empathetic samples to train on, or pre-training conversational data.

4.4.4 *Human Evaluation*

Automated metrics such as PPL and BLEU scores are not sufficient alone to evaluate a model's ability to exhibit empathetic behavior. Given the unavailability of specific metrics to evaluate empathy in a conversational model, we resort to evaluation based on the judgment of human subjects. Through human evaluation, we can evaluate the emotional communication capability of the models, which is their ability to recognize emotion in the input utterance and generate a suitable expression of emotion in their corresponding response [92]. To this end, we conducted a survey to collect ratings from 85 native Arabic speakers.

The raters were shown various utterances and their corresponding responses generated by the Baseline, EmoPrepend, and BERT2BERT models. The BERT2BERT-UN model was excluded from the survey given its poor results in terms of numerical metrics. The raters were asked to evaluate each of the models' ability to show Empathy, Relevance, and Fluency in the generated response. The raters were asked to answer the following questions:

- Empathy: Does the generated response show an ability to infer the emotions in the given utterance?
- Relevance: How relevant is the generated response to the input utterance?
- Fluency: How understandable is the generated response? Is it linguistically correct?

For each question, the raters were asked to score the responses of the models on a scale of 0 to 5, where 0 reflects extremely poor performance and 5 reflects excellent performance.

The results of the survey are summarized in Table 4.3, where we report the average of the collected ratings. The EmoPrepend model showed a higher average score of Empathy and Relevance than the Baseline model. However, these scores are below 3, meaning the EmoPrepend model was seen to deliver below-average performance.

On the other hand, the average ratings of the BERT2BERT model can be considered high and are much superior to both the Baseline and the EmoPrepend models,

Model	Empathy	Relevance	Fluency
Baseline [84]	2.04	1.68	2.44
EmoPrepend	2.81	2.18	3.28
BERT2BERT	**4.0**	**3.59**	**4.30**

Table 4.3: Average evaluation of the collected human ratings.

which indicates BERT2BERT's ability to deliver highly empathetic responses while abiding by linguistic correctness. This is reflected in some examples of the generated responses by BERT2BERT that can be seen in Table 4.4. The responses demonstrate the model's ability to express empathetic, relevant, and fluent responses when prompted with input utterances with various emotional states and domain contexts, which also proves its ability to handle open-domain conversations.

Generated Response	Utterance	Emotion
أوه لا! ماذا حدث ؟	اليوم يوم سيئ جدا	Sadness
!من الرائع سماع ذلك ما الذي يجعلك سعيدا جدا حيال ذلك ؟	أنا سعيد جدا	Joy
أنا آسف جدا لسماع ذلك. هل أنت بخير ؟	لقد أصبت بفيروس كورونا	Fear
إنه لأمر رائع أنك فخور جدا بنفسك	أنا فخور جدا بكوني مهندس	Joy
تهانينا! ما هو العمل الشاق ؟	لقد تلقيت ترقية في عملي اليوم	Surprise
أنا اسف لسماع ذلك هل كان عيد ميلادك المفضل ؟	خلال عيد الميلاد قبل بضع سنوات، لم أحصل على أي هدايا	Sadness
أوه ، واو! يجب أن يكون ذلك مزعجا حقا	يستمر جاري في إحداث ضوضاء في الساعة الثانية صباحا	Anger
أوه ، واو ، أنا متأكد من أنك ستكون بخير	أختي ستتزوج الأسبوع المقبل . أنا سعيد جدا لها ولكن في بعض الأحيان أشعر بشيء ثقيل في القلب	Sadness

Table 4.4: Examples of responses generated by the BERT2BERT model for multiple utterances with various emotional states and domain contexts.

4.4.5 *Performance on Inputs with Neutral Emotional States*

Despite the promising results achieved by the BERT2BERT model in generating relevant empathetic responses in open-domain settings, it was shown to poorly handle regular chit-chat utterances with neutral emotions, such as "Hey, how are you?" or "What are you doing?". Instead of providing a regular response, the BERT2BERT model will opt to generate an empathetic response as we show in Table 4.5. This issue can be explained by the fact that the model was fine-tuned on a dataset comprised of utterances with pure emotional context and corresponding empathetic responses. Moreso, the AraBERT-initialized parameters did not help mitigate this issue since AraBERT is pre-trained in a self-supervised fashion on news articles and later fine-tuned on a

task-specific dataset that does not contain regular chit-chat samples. Thus, it is clear why the BERT2BERT model is not able to handle neutral chit-chat conversations, as it is outside the scope of the training data and the task at hand.

Generated Response	Utterance
أنا سعيد لسماع ذلك	مرحبا كيف حالك ؟
ماذا عنك ؟	ماذا تفعل ؟
أوه ، واو. أنا أحب الحيوانات الأليفة	ما اسمك ؟
لست متأكدا من أين ذهبت. هل ذهبت ؟	هل زرت لاس فيجاس من قبل ؟
هذا يبدو و كأنه عطلة نهاية أسبوع رائعة	سأذهب للتنزه قليلا و أعود بعد قليل

Table 4.5: Examples of responses generated by the BERT2BERT model for multiple utterances with neutral emotions.

CHAPTER 5

DIALOGUE RESPONSE GENERATION IN ARABIC DIALETS WITH SELF-SUPERVISED LEARNING

5.1 Objectives

In low resource scenarios, such as language varieties and dialects, open-domain conversational data is limited and very expensive to obtain. Also, pre-trained and dialect-specific language generation models are unavailable. The lack of these resources motivates our research question: "*Can we leverage pre-trained language understanding models to develop open-domain response generation models in specific dialects of a language, where the needed resources are not sufficient?*" In this work, we aim to answer this question by warm-starting a transformer with pre-trained language model parameters, adapting the model to the target dialect via self-supervised pre-training on a large corpus of unlabeled dialectal samples, and finally fine-tuning the adapted model on a small number of open-domain dialectal conversational samples.

To study the effectiveness of the proposed approach, we apply it to Arabic as it offers a rich context of linguistic variations, and since the majority of the available resources are in Modern Standard Arabic (MSA), with very limited existing resources for individual dialects. Arabic has many spoken dialects [93]–[95], with significant differences to the written MSA on morphological, lexical, syntactic, and phonological levels. Not only do samples in Dialectal Arabic (DA) include slang, idioms, and similar expressions that deviate from the MSA norms, but they also do not follow any standard orthography [96], creating spurious data sparseness and making it much more challenging to process DA samples, especially for sequence generation tasks. The results show significant performance improvements on the three most widely spoken Arabic dialects after adopting the framework's three stages.

5.2 Related Work

low-resource-response-generation studied the problem of low-resource response generation with 360K utterance-response pairs in Chinese. The authors proposed estimating templates from large-scale unlabeled samples to aid an encoder-decoder model in response generation. naous achieved high performance in open-domain response generation in MSA by fine-tuning warm-started a transformer model on 36K utterance-response samples that were automatically translated from English **naous2**. In our work though, we tackle the problem of open-domain response generation in DA with only 1K utterance-response pairs. Specifically for DA, a few works have been proposed [19], [97], all of which are closed-domain that can handle a few specific topics. Additionally, these works rely on rule-based approaches or retrieval systems which limits their ability to generalize on unforeseen domains, as opposed to our generation-based approach which leverages knowledge from large-scale pre-trained language models. To the best of our knowledge, this is the first attempt to tackle open-domain response generation in DA.

5.3 Proposed Method

Consider the learning task $\mathcal{T}$ of generating an open domain response in a low-resource language, such as DA. Let D_{train} denote the dataset used to learn the task $\mathcal{T}$. D_{train} contains I samples of dialectal utterance-response pairs $\{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^{I}$, where $\mathbf{x} = [x_0, x_1, \ldots, x_{n_x}]$ represents an utterance of length n_x tokens and $\mathbf{y} = [y_0, y_1, \ldots, y_{n_y}]$ is the response of length n_y tokens. By fitting on D_{train}, the goal is to learn the weights θ that parameterize a hypothesis $h(.;\theta)$ that best approximates the optimal hypothesis h^* in a hypothesis space $\mathcal{H}$. However, given the low-resource setting and the insufficient amount of samples I, a reliable approximation of h^* cannot be reached by merely fitting the model on D_{train}. Our proposed approach, illustrated in Fig. 5.1, addresses this challenge and consists of three stages: warm-starting the encoder-decoder model, self-supervised pre-training on target dialect, and fine-tuning on target dialect task.

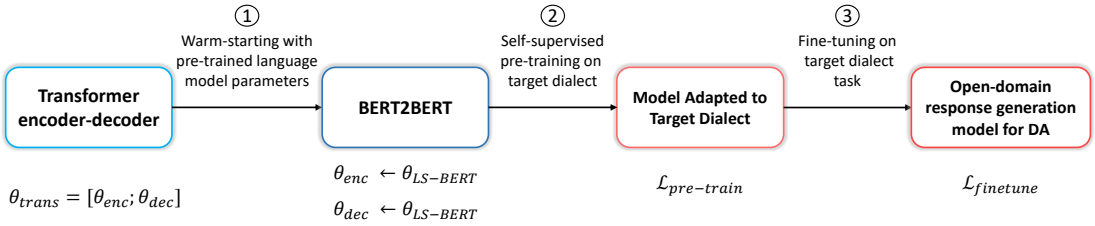

Figure 5.1: Illustration of the three stages (warm-starting, self-supervision, and pre-training) of the proposed framework for learning open-domain response generation in DA.

5.3.1 *Warm-Starting of Encoder-Decoder for the Standard Language*

At the first stage, we start with an encoder-decoder model and choose the transformer architecture [98]. We denote by θ_{trans} the parameters of the transformer model which consist of the encoder parameters θ_{enc} and of the decoder parameters θ_{dec}. The model is warm-started with pre-trained Language-Specific BERT (LS-BERT) parameters, an approach that has shown its ability in leveraging knowledge transfer from a Natural-language Understanding (NLU) model for sequence generation tasks **naous-bert2bert**, [83]. Specifically, the weight parameters of all layers of the encoder that also exist in the LS-BERT, such as self-attention and feed-forward layers, are initialized with LS-BERT parameters. The same procedure is done for the decoder:

$$\begin{aligned} \theta_{enc} &\leftarrow \theta_{LS-BERT} \\ \theta_{dec} &\leftarrow \theta_{LS-BERT} \end{aligned} \tag{5.1}$$

Hence, warm-starting the transformer will result in an encoder-initialized encoder and an encoder-initialized decoder. For layers that do not co-exist in both the transformer encoder-decoder model and the LS-BERT, specifically cross-attention layers, weights are randomly initialized. For Arabic, we experiment with three different language models: AraBERT [18], ARBERT [99], and MARBERT [99]. AraBERT and ARBERT have been pre-trained on MSA samples. Warm-starting the transformers with the parameters of these models helps leverage lexical similarity between MSA and DA and transfer knowledge of such words. MARBERT has been pre-trained on both MSA and DA samples, which helps acquire prior knowledge of certain dialectal sub-tokens.

5.3.2 *Self-Supervised Pre-training on Target Dialect*

In the second stage, the warm-started model is pre-trained in a self-supervised manner on a large number J of unlabeled data dialectal sequences, denoted by $D_{pre-train} = \{\mathbf{x_j}\}_{j=1}^{J}$, that is closer to the distribution of the target task $\mathcal{T}$. We adopt the next sentence pre-training strategy, an extension of the GPT causal language modeling strategy for encoder-decoder models. In this setup, the model is trained to generate the next part of a DA sentence. This pre-training step adapts the model to the target dialect and learn semantic and syntactic information it needs to generate fluent DA sequences.

5.3.3 *Fine-tuning on Dialectal Response Generation Task*

At the last stage, we fine-tune the model on D_{train} where the negative log-likelihood of the target dialectal response y given the input utterance x in the data distribution $P_{D_{train}}$ is minimized.

$$\begin{aligned}\mathcal{L}_{finetune} &= E_{(x,y)\sim P_{D_{train}}}(-\log P_\theta(\mathbf{y}|\mathbf{x})) \\ &= E_{(x,y)\sim P_{D_{train}}}(-\log \prod_{t=1}^{n_y} P_\theta(y_t|y_{<t}, \mathbf{x}))\end{aligned} \tag{5.2}$$

where $y_{<t}$ represents $[y_0, y_1, \ldots, y_{t-1}]$.

5.4 Datasets

5.4.1 *Twitter Corpora Collection for Self-Supervised Pre-training*

Collection: We collected ∼1M tweets in each dialect to perform self-supervised pre-training. To ensure that the tweets collected are specifically in the Levantine, Egyptian, and Gulf dialects and do not contain samples written in MSA, the Twitter accounts were manually inspected. In each dialect, 500 Twitter accounts were selected for scraping using Tweepy after human inspection. These accounts mainly belong to native speakers of each area in the Arab world (LEV, EGY, GUL) with various educational and societal backgrounds where the tweets offer their personal opinions on various topics.

Pre-processing: To have clean and meaningful sentences for pre-training, the tweets were pre-processed to remove English sentences, duplicate tweets, non-Arabic tokens, hashtags, symbols, numbers, and emojis. We also removed tweets that hold a length below 5 after tokenization. The tweets are then split in half for self-supervised pre-training with the next sentence generation objective.

5.4.2 *Message-Response Pairs Datasets for Fine-Tuning*

We hired three native translators from the Levantine, Egyptian, and Gulf areas of the Arab world to translate 1K samples from the DialyDialog dataset to their specific dialects. DailyDialog [100] is a high-quality multi-turn dataset of ∼12K open-domain dialogues covering ∼1.4M English words and is commonly used in the English literature on open-domain response generation. Hence, we compiled 1K open-domain message-response samples in each dialect that we used to fine-tune our model in the last step of the framework.

5.4.3 *Vocabulary Overlap*

We report in Table 5.1 the vocabulary overlap between the collected, pre-processed tweets and the translated message-response pairs in each dialect. The results are reported with and without the use of Farasa segmentation [80]. When using Farasa segmentation, we notice that a very large overlap (¿90% for all dialects) exists between the corpora in each dialect. This overlap is unsurprisingly reduced when Farasa is

not used but is still considered good overlap (¿70% for LEV and EGY, and ¿50% for GUL). Thus, the self-supervised pre-training of the model on the unlabeled samples is expected to boost the performance of the model, as it would gain prior knowledge of these tokens before being fine-tuned for response generation.

(a) With Farasa Segmentation				
		Tweets		
		LEV	**EGY**	**GUL**
Translated Message-Response Samples	**LEV**	90.26%		
	EGY		90.12%	
	GUL			91.25%
(b) Without Farasa Segmentation				
		Tweets		
		LEV	**EGY**	**GUL**
Translated Message-Response Samples	**LEV**	77.30%		
	EGY		76.75%	
	GUL			54.63%

Table 5.1: Vocabulary overlap between the translated samples and the scraped tweets in each dialect with **(a)** and without **(b)** Farasa Segmentation.

5.5 Experiments & Results

5.5.1 *Experimental Setup*

Datasets: We evaluate our framework on the three most widely spoken Arabic dialects: Levantine **(LEV)**, Egyptian **(EGY)**, and Gulf **(GUL)**. To perform self-supervised pre-training, we resort to Twitter as a resource that is rich in DA text. For each dialect, we scraped ∼1M tweets using Tweepy[1]. For fine-tuning, we hired native translators for each dialect to create small datasets of open-domain dialectal utterances-responses pairs in three dialects. To ensure the quality of the dataset, we randomly sampled 1K dialogues from the DailyDialog dataset [100] and asked the translators to write equivalent utterance-response pairs in their dialect, with modifications where necessary (change of English names or references such as location, etc.).

Evaluation Metrics: We perform evaluations using both automated metrics and human judgment. For automated metrics, we use the **Perplexity** (PPL) and **BLEU** scores which are commonly used for the evaluation of natural language generation models. For human evaluation, we ask three well educated and native speakers in each dialect to rate the responses (on a scale from 1 to 5) of the test set samples from three aspects: (1) **Fluency:** if the response is a meaningful sentence that does not

[1]https://docs.tweepy.org

contain grammatical errors, (2) **Relevance:** if the response is relevant to the topic in the input message, and (3) **Richness:** if the response contains interesting and new content. Details on the human evaluation process are provided in Appendix **??**.

Implementation: We base our code on the Huggingface transformers library [101]. The checkpoints of the AraBERT, ARBERT, and MARBERT language models are used to warm-start the transformer model. The Adam optimization algorithm is used for all experiments. The fine-tuning dataset is split into 80% training, 10% validation, and 10% testing.

Baselines: We compare our proposed approach to the following baselines: (1) **Seq2Seq** [102]: the basic sequence-to-sequence recurrent neural network model with Long Short-Term Memory (LSTM) units and Attention, (2) **Transformer** [98]: the transformer sequence-to-sequence architecture that is based on attention mechanisms, and (3) **BERT2BERT** [84]: the transformer architecture with encoder and decoder layers warm-started with pre-trained BERT parameters.

5.5.2 *Evaluation Results*

Automatic Evaluation: The results in terms of automated metrics on the test set are reported in Table 5.2. Training a regular Seq2Seq or Transformer model from scratch on the utterance-response datasets yields very poor performance on all dialects due to the small number of samples (1K in each dialect). The addition of the step of Self-Supervised Pre-training (SSP) using the pre-processed tweets before fine-tuning (FT) the transformer model on response generation shows improvements on both PPL and BLEU scores but is not enough alone for good performance since the scores are still considered poor. Warm-starting the transformer model using the pre-trained parameters of Arabic language models shows great improvements prior to performing SSP and fine-tuning. Compared with regular fine-tuning of warm-started transformers (BERT2BERT models), the step of SSP with warm-starting before fine-tuning shows improvements on automated metrics for each dialect in the majority of the cases. The obtained results vary when using different pre-trained language models for warm-starting, depending on what data the BERT models were trained on and the vocabulary overlap with our message-response samples, which is affected by the type of segmentation used. For instance, AraBERT uses Farasa segmentation [80] while ARBERT and MARBERT do not. It is very clear though, from all experiments, that adding SSP and FT improves the results regardless of the dialect.

Pre-trained Language Model	Model	LEV		EGY		GUL	
		PPL (↓)	BLEU (↑)	PPL (↓)	BLEU (↑)	PPL (↓)	BLEU (↑)
×	Seq2seq	329.57	0.001	383.51	0.001	284.18	0.027
×	Transformer	283.14	0.001	354.15	0.001	269.02	0.001
×	Transformer + SSP + FT	211.86	0.284	294.18	0.137	168.53	0.202
AraBERT	BERT2BERT	231.59	0.216	260.03	0.127	91.46	0.665
	BERT2BERT + SSP + FT	65.73	**1.234**	376.73	0.028	52.36	1.416
ArBERT	BERT2BERT	520.55	0.441	470.91	0.255	168.98	1.191
	BERT2BERT + SSP + FT	321.82	0.976	435.58	0.294	83.15	**1.831**
MARBERT	BERT2BERT	487.27	0.629	528.68	0.640	343.52	0.320
	BERT2BERT + SSP + FT	358.49	0.723	389.61	**0.973**	317.15	0.648

Table 5.2: Automatic evaluation results on the test set. SSP stands for Self-Supervised Pre-training. FT stands for Fine-tuning. The results show improvements on both automatic metrics in the majority of the cases when the three stages of the framework are used: Warm Starting (BERT2BERT) followed by SSP then FT. We note that the models cannot be directly compared in terms of PPL due to the usage of different segmentations in pre-trained language models. Bolded numbers indicated the highest achieved BLEU score in each dialect.

Human Evaluation We generated 100 responses using the test set in each dialect. For each dialect, we invited three well-educated native speakers of the dialect to rate each response in terms of Fluency, Relevance, and Richness. The raters were given the rating key shown in Table 5.3. The results were then averaged to report the overall ratings in each dialect (Table 5.4).

Fluency	
Rating	*Description*
0	The response is full of grammatical mistakes and is incomprehensible
3	The response contains one or two grammatical mistakes but is somewhat comprehensible
5	The response contains no grammatical mistakes and is very comperehsible
Relevance	
Rating	*Description*
0	The response is completely irrelevant to the message and hence is off-topic
3	The response is not directly relevant to the topic of the message but captures some detail
5	The response is totally relevant to the topic of the message
Richness	
Rating	*Description*
0	The response is poor in content (repetitive words and short or yes/no answers)
3	The response contains some new content (few words related to the topic message)
5	The response contains informative content and may keep a conversation going (personal opinion, questions)

Table 5.3: Human Evaluation Rating Key

We select the models that achieve the highest BLEU score in each dialect and generate responses for the 100 samples of the test set using the top-k sampling algorithm with k set to 50. The averaged ratings of the human evaluators are reported in Table 5.4. These results indicate high Fluency and Richness scores on the three dialects, which means that the models are able to generate coherent and meaningful sentences in each dialect with interesting content. However, lower scores on Relevance are obtained,

which indicates that the models can often generate off-topic responses. In addition to the very small number of utterance-response pairs used in fine-tuning, poor relevance can also be heavily influenced by the used decoding algorithm where k should be tuned to reach a reasonable output.

Dialect	Model	Fluency	Relevance	Richness
LEV	BERT2BERT (W.S. w/ AraBERT) + SSP + FT	4.276	1.467	4.213
EGY	BERT2BERT (W.S. w/ MARBERT) + SSP + FT	4.165	0.864	3.948
GUL	BERT2BERT (W.S. w/ ARBERT) + SSP + FT	4.543	1.354	4.158

Table 5.4: Averaged human evaluation scores for each Arabic dialect. W.S. stands for Warm-Started. Judgment was done on responses generated using top-k sampling with $k = 50$. The results indicate high fluency and richness in the responses but lower relevance.

5.5.3 *Examples Responses*

Cherry-Picked Examples: We show some cherry-picked responses from the test set in Table A.1 that have been generated using the models achieving the highest BLEU scores in each dialect as reported in Table 5.2. The responses were generated using top-k sampling with k set to 50. In these examples, the model can understand the topic in the given message and generate a relevant response without any grammatical errors. Ideally, the response would be constructed in a way that is relevant to the message, contains specific information, and can keep a conversation with the user going. This is done for example in cases where questions are asked in the response (e.g., asking why there are problems or where the user is going) of the model which will more likely keep the user engaged compared with responses that only deliver relevant content (e.g., congratulating the user on winning the lottery).

Lemon-Picked Examples We show some lemon-picked responses from the test set in Table A.2 that have been generated using the models achieving the highest BLEU scores in each dialect as reported in Table 5.2. The responses were generated using top-k sampling with k set to 50. These results expose various errors that the models can face. First, the model can understand the context of the message but responds in an incorrect manner such as in the example in LEV where the model responds Saturday evening whereas the message was asking about tomorrow morning. Second, the model may provide a relevant response but would contain a grammatical error such as in the example in GUL (I went I have an office). However, grammatical mistakes are influenced by the value of k used in sampling and can be mitigated by lowering k. This would help reduce these types of errors at the expense of having more repetitive and boresome generations, which is a drawback of the common neural sampling algorithms where k needs to be tuned [103]. One area of research targeting this problem is perplexity-controlled sampling algorithms [104]. Finally, in the worst cases, the

model would provide completely irrelevant responses which have been most noticeable in EGY samples.

5.6 Ethical Considerations

The pre-training stage of our models is based on data collected from Twitter and which could contain toxic, offensive, and biased content since they have been produced by third-party Internet users. The pre-trained language models used for warm-starting may also suffer from similar problems, especially the ones that contain DA content in their pre-training. However, our fine-tuning stage is based on manually translated DA samples by hired translators and does not contain any offensive or biased content, which can mitigate any learned unethical behavior during warm starting and self-supervised pre-training. Upon inspection of test-set generated responses, we also did not notice any of these issues. However, these problems can be reduced in the pre-training stage of the framework by using safety classifiers to filter out Twitter-collected samples containing toxic content. Developing such classifiers for DA hate speech and offensive language detection is an ongoing area of research [105], [106].

CHAPTER 6

RETRIEVAL-REINFORCED MAXIMUM SIMILARITY DECODING

6.1 Motivation and Objectives

The quality of output generated by text generation models is strongly influenced by neural decoding algorithms. The most widely adopted decoding algorithms such as top-k and top-p sampling with temperature require an ad-hoc parameter tuning to reach satisfactory performance. It has been empirically observed that lowering k and p yields coherent yet repetitive generations, while increasing k and p produces more surprising and rich content at the risk of reduced coherence of the text. Further, in the case of dialogue response generation, one parameter configuration that produces relevant output for one example would not produce an irrelevant output for a different example.

MSG	I forgot my wallet at home this morning!
RSP (p=0.7;t=0.9)	Oh no! were you able to keep it ?
RSP (p=0.8;t=0.7)	That's so embarrassing! did you go back home?
MSG	I just got promoted at work today!
RSP (p=0.8;t=0.7)	Congratulations! are you alright ?
RSP (p=0.9;t=0.4)	That's awesome! did you get a raise as well ?

Table 6.1: Responses generated by BERT2BERT fine-tuned for empathetic response generation using different parameter configurations for top-p sampling with temperature (t). We observe that one configuration for decoding hyper-parameters cannot do well for all input messages. MSG stands for Message. RSP stands for generated response.

We show this problem in Table 6.1 where a BERT2BERT model fine-tuned for empathetic dialogue response generation is used to generate responses for different configurations of top-p sampling with temperature. After manually finding a hyper-parameter configuration that produces the most relevant response for the first input

message, using the same configuration to decode a response for another message does not perform as well. In fact, using a different parameter configuration for the second input result in a more relevant response. This motivates the need of a method that tunes the sampling-based decoding technique's hyper-parameters per input message to generate the most relevant response possible.

In this work, we address this issue by proposing an algorithm that searches for the optimal hyper-parameters of the decoding technique that provide the most relevant response. We propose to use a retrieval-reinforced approach where the algorithm searches for the response that maximizes a distance measure to the most likely retrieved candidates. We verify the effectiveness of the approach by testing it on both English and Arabic datasets for open-domain dialogue and comparing its results to the ones obtained by having fixed choices of hyper-parameters.

6.2 Related Work

Open-domain Dialogue Models: Recent work on open-domain dialogue can be classified into two different approaches: retrieval-based approaches which select the most likely response from a fixed set of dialogues, and generation-based approaches that decode responses from a learned model distribution. The success in both these approaches has been achieved by adopting neural pre-trained models that require training on a large quantity of high-quality data [107]. While retrieval-based systems can guarantee correct grammar and accurate information, they can produce unsatisfactory responses when the context is substantially different from the dialogues available to retrieve from. On the other hand, generation-based systems implicitly store knowledge in the model's parameters, which allows to generate novel responses from scratch. Still, generative models can produce dull responses, inaccurate information, and poor grammar [108].

Neural Decoding Algorithms: The performance of generative models is heavily influence by the mechanism in which the output is decoded. Deterministic approaches such as greedy search or beam search and its variants tend to produce repetitive responses, which results in a boring human-machine interaction experience [109]. Introducing randomness through sampling into the decoding process mitigates this problem and allows producing more surprising responses. The most widely adopted sampling-based decoding techniques are top-k and top-p sampling. In top-k sampling, the probability mass is redistributed among the k most likely tokens, while top-p samples from the smallest number of tokens of which the cumulative probability exceeds p, avoiding problems of very shap or flat distributions obtained in top-k. The temperature t parameter is also often introduced in the softmax operation to control the sharpness of the distribution. However, the selection of those parameter is done in an ad-hoc manner since there is no clear way to select them. A recent attempt in addressing this problem [104] for text generation proposes an adaptive top-k sampling approach that tunes k according to a target desired surprise value. Our work addresses the tuning of

sampling-based decoding parameters in a one-to-many setting with focus on the task of dialogue response generation where many outputs could be reasonable.

Retrieval-Augmented Generation: A new emerging paradigm in text generation is retrieval-augmented generation methods that aim at addressing the problems faced with generation-based methods by fusing them with retrieval-based methods [110]. This new paradigm offers advantages over conventional generation, such as the ability to acquire information from external sources instead of relying solely on the information learned via training, offering better scalability [111]. Retrieval-augmented generation methods have been used for dialogue generation [112]–[116], text summarization [117], [118], and other text-to-text generation applications [119]–[121]. Those works however ignore the influence of the decoding mechanism on the generated output. Our work leverages retrieval to tune the sampling-based decoding parameters per input context.

6.3 Method

We propose RRMSD[1], a decoding strategy to enhance relevance in dialogue response generation by leveraging most similar candidates from the training data.

6.3.1 *Core Algorithm*

The algorithm pseudo-code is presented in Algorithm 1. First, we use a pre-trained sentence encoder $\mu_{\theta^*}(.)$ that converts the training data D into vector representations V_D. We are interested in retrieving input sentences from D that are most semantically similar to the input of the algorithm. For an input message x_{test}, we measure its semantic textual similarity to every input message $x_i \in D_{i=1}^{m}$. This is performed by computing a distance measure (cosine similarity s) between x_{test} and every x_i in the sentence encoder's embedding space. The training set samples are then ordered by descending order of similarity. We select the n most similar sentences x_{α_i} from the sorted samples and retrieve their corresponding ground-truth responses y_{α_i} where $i \in \{1, n\}$.

Next, a brute force search over the parameter ranges of top-p sampling with temperature (t) is performed. At each step of the search, we use the fine-tuned response generator model $h_\theta(.)$ in inference mode to generate a response $\hat{y}_{p;t}$ where the combination p and t of the search step is used as the choice of decoding hyper-parameters. The total similarity between $\hat{y}_{p;t}$ and y_{α_i} for $i \in \{1, n\}$ is computed, where both are encoded by the sentence encoder and the cosine similarity is used as the distance measure. We search for the response that achieves the maximum summed similarity with the n nearest responses retrieved in the first step, and which would be selected by the algorithm as the optimal model output that provides a relevant response to x_{test}.

[1]RRMSD: **R**etrieval-**R**einforced **M**aximum **S**imilarity **D**ecoding

Algorithm 1: Retrieval-Reinforced Maximum Similarity Decoding

Given : input message x_{test}, training set $D = \{x_i; y_i\}_{i=1}^{m}$, response generator $h_\theta(.)$, sentence encoder $\mu_{\theta^*}(.)$, and number of most similar messages n (hyperparameter) in D

$v_{test} = \mu_{\theta^*}(x_{test})$

$V_D = \{v_i; y_i\}_{i=1}^{m} = \{\mu_{\theta^*}(x_i); \mu_{\theta^*}(y_i)\})_{i=1}^{m}$

for $x_i \in D$ **do**

  $s_i = \frac{v_{test}.v_i}{||v_{test}||_2||v_i||_2}$

end

Select n y_i's from D that correspond to x_i's with largest similarities s_i with indices $(\alpha_1, \ldots, \alpha_n)$:

$R = \{r_i\}_{i=1}^{n} = [y_{\alpha_1}, \ldots, \alpha_n]$

for $p = 0.1;\ p <= 1;\ p = p + 0.1$ **do**

  for $t = 0.1;\ t <= 3;\ t = t + 0.1$ **do**

    $\hat{y}_{p;t} = h_\theta(x_{test}; p, t)$

    $s_{p,t} = \sum_{i=1}^{n} \frac{\hat{y}.r_i}{||\hat{y}_{test}||_2||r_i||_2}$

  end

end

Return $\hat{y}$ that result in the largest total similarity score $s_{p,t}$

6.3.2 *Choice of Sentence Encoder*

The choice of sentence encoder is a critical part of the algorithm as it needs to map inputs into an embedding space that allows us to efficiently find sentences that are semantically similar. Using a regular pre-trained encoder model (e.g. BERT) would not be sufficient because **1)** instead of sentence embeddings, they provide embedding vectors per token which would need to be mean pooled for use in measuring distance to other sentences, and **2)** their construction require large inference time which makes them not scalable to use for retrieval on datasets with tens of thousands of samples. Instead, we use sentence transformers, which are fine-tuned on natural language inference tasks to learn semantically meaningful sentence embeddings, and which provide huge computational advantages over regular encoders [122]. Specifically, we select the MPNet [123] sentence transformer that is trained on more than 1 billion training pairs.

6.4 Experimental Results

6.4.1 *Automatic Evaluation*

Table 6.2 reports the results of RRMSD on the test set of the Arabic-translated version of the Empathetic Dialogues dataset in both English and Arabic versions, compared

to the results obtained by using fixed parameter configurations of top-p sampling with temperature. Table 6.3 shows a similar comparison when RRMSD is used to decode responses on the English Empathetic Dialogues using a BERT2BERT model fine-tuned on the dataset. The comparisons are done using two metrics: the BLEU score, and the Semantic Textual Similarity (STS) score which we compute using the cosine similarity on sentence transformer embeddings. The results clearly show the enhancements provided by RRMSD on all metrics, and in both languages. This indicates that overall, RRMSD can help mitigate cases where the generated response by the model goes off topic, thus providing outputs which are more semantically relevant to the ground-truth data.

Decoding Algorithm		*Arabic*	
		STS	BLEU
RRMSD		**0.1887**	**0.2024**
top-p sampling	*p=0.7 ; t=1.5*	0.1546	0.1915
with temperature	*p=0.9 ; t=0.8*	0.1497	0.1873

Table 6.2: Results achieved by RRMSD on the Arabic Empathetic Dialogues dataset. STS stands for Semantic Textual Similarity measured using the cosine similarity on sentence transformer embeddings.

Decoding Algorithm		*English*	
		STS	BLEU
RRMSD		**0.2561**	**0.139**
top-p sampling	*p=0.7 ; t=1.5*	0.1832	0.12
with temperature	*p=0.9 ; t=0.8*	0.195	0.125

Table 6.3: Results achieved by RRMSD on the Empathetic Dialogues dataset. STS stands for Semantic Textual Similarity measured using the cosine similarity on sentence transformer embeddings.

6.4.2 *Example Responses*

Table 6.4 shows some responses decoded by RRMSD in Arabic and compares them with the responses obtained through vanilla top-p with temperature where p is set to 0.7 and t is set to 1.4. It can be clearly observed in those samples that the quality obtained through RRMSD is much better and provides better relevance to the input message. For example, in the first example where the message signals that the user's daughter is suffering from an infection, the output obtained through regular sampling asks the user if she will take her daughter to the vet, confusing her with a certain animal such as a cat. On the other hand, RRMSD avoided such a situation and she asked if the user thinks her daughter will be alright. This is because the output obtained by

regular sampling would likely have a low similarity score with the retrieved response candidates where the reference to a veterinarian should not appear.

top-p sampling with temperature (p=0.7; t=1.4)	RRMSD	Input
يا هذا مروع ، هل كان علي كأن تأخذها إلى الطبيب البيطري ؟	أوه لا! هل تعتقد أنها بخير ؟	ابنتي الصغرى تعاني من الحساسية
واو هذا مقرف ، آسف لسماع ذلك	أوه لا ، هذا فظيع. ماذا فعلت ؟	استيقظت هذا الصباح ووجدت شاحنتي مثقوبة في الإطارات
هل أنت مريض أم أنه سيكون على ما يرام ؟	أوه ، آمل أن تكون على ما يرام! هل أنت بخير الآن ؟	إنه حقًا صعب ، لقد بدأت العلاج مؤخرًا
!آه حسنا أتمنى أن أعود إلى المدرسة	هل وجدت شيئا ؟	كنت أبحث في الصور القديمة لأطفالي الصغار
أوه لا ، أتمنى أن يكون الأسوأ آمل ألا يكون الأمر مخيفا	أوه لا! هذا مقرف. هل لديك أي سبب وجيه ؟	مرحبا صديقي. لقد استلمت للتو نتائج ترقيتي واتضح أنني لم يتم اختياري - وهو ما يمثل مشكلة هائلة
ذلك رائع! أراهن أنه شعر بشعور جيد عندما سمعت ذلك	هذا رائع! كم عمره ؟	!قال طفلي ماما لأول مرة ... لم أصدق ذلك
ربما تحتاج إلى تغيير نظامها	هل تعتقد أنك ستنتقل أم لا ؟	أنا قلق حقًا من أن تهاجمنا كوريا الشمالية

Table 6.4: Example generated results by RRMSD compared to choosing a fixed hyper-parameter configuration of top-p sampling with temperature.

6.4.3 *Computational Time*

While the proposed algorithm provides enhanced relevance in the generated responses, it suffers from a larger computational time since it needs to search over a wide range of sampling parameter configurations. However, this time is constant since the combinations are fixed and the inference time inside the loop is also fixed but is a subject of the hardware it is running on. Using an NVIDIA Tesla K80, the algorithm takes around 90 seconds to provide a response to one input, which is very large. Running the algorithm on more powerful GPUs (NVIDIA P100) leads to a decrease in its running time to around 60 seconds, but is still considered high. Ultimately, we would need this running time to be below the 5 seconds margin to make it applicable for usage in a real-time chatbot application, because users cannot wait for long periods of time for a response as it will make the human-computer interaction experience a boring one.

As future directions, we aim to mitigate computational time by reducing the search space of the algorithm while still obtaining similar performance. Specifically, we aim to analyze which parts of the distributions would it make sense to search for the maximum similarity. For example, decoding using small probability values (e.g. 0.1, 0.2, or 0.3) would make the algorithm greedy and usually results in almost the same responses even when temperature is varied. In such cases, the algorithm would be wasting time considering the same outputs for similarity. Additionally, in such small probability values for decoding, the response is usually brief and generic which we want to avoid. It would make better sense not to search when using such probabilities, but rather when using higher probabilities, where more tokens are considered and the more rich-in-content and relevant response could be generated at the expense of more risk of going-off-topic and introducing irrelevant words.

CHAPTER 7

CONCLUSION

Open-domain dialog agents are systems that are expected to engage coherently and engagingly in conversations with human users. Being ”open-domain”, such types of agents are not restricted to specific domains such as medical advice, customer support, or other industrial setting, where agents aim to achieve a specific goal that is serving the user. Rather, open-domain dialog systems have a more open-ended goal where they are required to converse on any potential topic in a fluent human-like language. Developing such types of agents could offer huge enhancements to the human-computer interaction experience. While this objective is a challenging one, the recent advances in deep learning and natural language processing applied to conversational artificial intelligence have resulted in exciting and promising results in open-domain dialog. Recent attempts have leveraged large-scale unlabeled data to develop pre-trained text generation models, which are then fine-tuned on crowd-sourced open-domain conversations, or used in a zero-shot manner to produce responses. While such approaches have shown success for resource rich languages such as English, low-resource languages such as Arabic and its dialects that do not have such pre-trained models or dialog datasets still lag behind.

This thesis targets the development of open-domain conversational bots in low-resource settings with a focus on the Arabic language and its dialects. Previous work in Arabic conversational AI proposed rule-based or retrieval-based approaches that are designed specifically for special applications and that cannot generalize to any domain. The main challenges in developing open-domain dialogue response generation models in Arabic is the lack of resources in terms of pre-trained language generation models and datasets of message-response pairs for training. This problem is augmented with dealing with Dialectal Arabic (DA) which suffers from a more severe lack of resources compared with Modern Standard Arabic (MSA).

In this thesis, we addressed those low-resource challenges and achieved state-of-the-art performance on dialogue response generation for both MSA and DA. First, we created a dataset in MSA of empathetic messages-response pairs by automatically translating samples from an English dataset for empathetic dialogue generation. The dataset was used to train a recurrent neural network model for sequence-to-sequence

generation and achieved average performance, with human judgement indicating that the model often does not produce relevant responses although the output is fluent in terms of language. Training a model from scratch on the message-response pairs was not enough since this task-specific dataset is considered small in size (38K samples). With the lack of pre-trained model for text generation that can be fine-tuned directly, we adopted transfer learning strategies for response generation by leveraging pre-trained checkpoints of a natural language understanding model (AraBERT) to initialize the parameters of a transformer-based encoder-decoder model. This model, named BERT2BERT, was then fine-tuned on our task of open-domain empathetic dialogue response generation and achieved state-of-the-art performance in MSA on automated metrics, which was also validated by human evaluation that confirmed the ability of the model to provide topic and emotion-relevant responses to the user's queries.

While the previous BERT2BERT model works well for MSA, it does not work well for DA since the AraBERT model used for initialization was predominantly pre-trained on MSA data. Additionally, open-domain message-response pairs in specific dialects were not available for fine-tuning. To extend the work done in MSA to DA, we proposed a three-stage learning framework based on warm-starting, self-supervised pre-training, and fine-tuning, which produced models that generate fluent response in DA. The approach first starts by initializing the encoder-decoder model by pre-trained checkpoints of BERT-based models, a step that helps in leveraging lexical similarity between MSA and DA. The model was then adapted to specific dialects by self-supervised pre-training on large-scale unlabeled data in the desired dialect. For self-supervised pre-training, 1M tweets in three dialects (Levantine, Egyptian, and Gulf) were scraped. For the last step of few-shot fine-tuning, small message-response pairs datasets in those dialects were manually crafted using the help of native speakers from those areas. The three-stage approach was implemented and tested using a variety of BERT models for initialization and showed enhanced performance in open-domain response generation in DA compared with multiple baselines. Specifically, models were able to generate coherent responses in multiple dialects, but still suffered from poor relevance.

Finally, the last part of thesis presents a decoding algorithm targeted at enhancing relevance in open-domain dialogue response generation. It was empirically observed that the relevance and content-richness of the response generated by such type of models are heavily influence by the choice of sampling parameters used. The most popular algorithms for decoding are top-k and top-p sampling with temperature, where the parameters of those algorithms are selected to a fixed configuration in an ad-hoc manner. However, we noticed that for one configuration that offers an optimal response for one input message, it does the opposite for a different message where a different parameter configuration provides an optimal response for it. This motivated the creation of a decoding algorithm that searches for the optimal parameters at each input message to provide the optimally relevant response. The approaches leverages a state-of-the-art sentence transformer model that measures semantic textual similarity to retrieve the responses of most similar candidates in the training set. The algorithm then searches in

a brute-force manner through different sampling parameter configurations to find the response that maximizes a distance measure (cosine similarity) to the retrieved candidates. This retrieval-reinforced maximum similarity decoding algorithm showed very promising results on automated metrics, indicating that it can provide overall more relevant responses, which was also observed through human error analysis.

Future work will expand on the developed decoding algorithm to further showcase its wide applicability by testing it on a wide range of open-domain dialogue datasets in English, and by using a variety of pre-trained state-of-the-art language generation models. In addition, the models will be evaluated using more automated metrics that measure various aspects of performance such as diversity. We also plan to design a human-evaluation scheme to assess and compare the performance of the models based on human judgement from various aspects. However, before such evaluations can be made on test sets with thousands of example, the running time of the algorithm needs to be decreased significantly to a reasonable level. Further, the algorithm in its current form does not consider the fact that at each run of the sampling algorithm, a different response is obtained, especially when sampling is being done from a larger vocabulary. We aim to incorporate this aspect in the algorithm in future work as it can capture more interesting responses that might be missed when just performing a single run.

Additionally, the datasets developed in this thesis consist only of a single message and its corresponding response per sample. When using the models trained on those dataset for a chatbot application, it will be difficult for the bot to make sense of the whole conversation history since it will only be conditioned on the previous user utterance to generate the response (in a single-turn manner). To overcome this, future work will focus on creating a high-quality multi-turn dataset that truly encapsulates human dialogue rather than relying only on a single turn-based dataset. Such type of dataset is difficult to obtain through manual crafting or automatic translation and would need to be gathered via crowd-sourcing, which presents challenges of its own in under-resourced languages.

APPENDIX A

ABBREVIATIONS

BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
DA	Dialectal Arabic
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
MSA	Modern Standard Arabic
NLG	Natural Language Generation
NLU	Natural Language Understanding
PPL	Perplexity
Seq2Seq	Sequence-to-Sequence

A.1 Example Responses in Dialectal Arabic

LEV	**Message:**	I won the lottery I'm gonna be rich! !ربحت اللوتو لح صير غني
	Generated Response:	Alright congratulations on winning ماشي الحال مبروك الفوز عليكن
	Message:	This town of full of problems that never end هالبلدة كلها مشاكل ما بتخلص
	Generated Response:	I'm not following any news, why ? I'm sure there's a lot of stuff مش عم تابع شي، ليش ؟ أكيد في اشيا كتير
	Message:	You look quite happy مبيين عليك مبسوط
	Generated Response:	Thank God I'm watching a nice movie الحمدلله بعدكم عم بحضر فيلم حلو
EGY	**Message:**	This country is full of problems that never end البلد ديه كلها مشاكل ما بتخلصش
	Generated Response:	What happened with you ? Don't worry! The important is that we eat ليه الي حاصل معاك ؟ مين يهمك! المهم حنا ناكل
	Message:	What's this beautiful dress that you're wearing ? ايه الطقم الحلو الي انت لابسه ده
	Generated Response:	Really I felt it looks fabulous on me بجد حسيته يجنن عليا
	Message:	Today you're invited to our evening gathering انهردة حضرتك معزوم على السهرة
	Generated Response:	No I'm not in the mood for anyone. Where are you going ? لا مش طايق حد. ليه رايح فين ؟
GUL	**Message:**	Can you study with the radio on ? تقدر تذاكر والراديو شغَّال؟
	Generated Response:	I can't. I can't. What do I do ? ما اقدر. ما اقدر. ايش اسوي ؟
	Message:	Hey, I'm David هلا، معاك داود
	Generated Response:	Hello! I saw you once here اهلا! شفتك مرة هنا
	Message:	Where did you go yesterday ? فين رحت أمس؟
	Generated Response:	I was waiting for the appointment so I went to the park جالس انتظر الموعد فرحت الحديقة

Table A.1: Cherry-picked examples generated using top-k sampling with $k = 50$.

LEV	**Message:**	We bought a new cat today اشتريتنا بسية جديدة اليوم
	Generated Response:	She's really nice don't rely on her a lot كتير لطيفة ما تتكل عليها كتير
	Message:	Can we go have breakfast together tomorrow morning ? ممكن نروح نتروق سوا بكرا ؟
	Generated Response:	Sure Saturday evening. Saturday is the best أكيد السبت بالليل. السبت احلى شي
	Message:	Can you study with music on ? في كتدرس أنت ومدور أغاني ؟
	Generated Response:	Good luck because I won the artificial intelligence contest بالتوفيق لأنو ربحت سباق الذكاء الاصطناعي
EGY	**Message:**	I really wanna travel! !قد ايه نفسي اسافر
	Generated Response:	No thanks! we need to go ? لا شكرا! لازم نروح ؟
	Message:	You're lying to me as well ? بتكذب عليا كمان ؟
	Generated Response:	No I don't wanna see. What happened with you ? لا مش عايز اشوف. ايه الي حاصل معاك ؟
	Message:	Where are you going to watch el classico ? حتتفرج على الكلاسيكو فين؟
	Generated Response:	There's nothing specific at all مفيش حاجة محددة خالص
GUL	**Message:**	Every nation should face history كل دولة لازم تواجه التاريخ
	Generated Response:	What happened 30 years ago and how much exactly ? ايش صار قبل ٣٠ سنة وكم بالظبط ؟
	Message:	Are you free today ? أنت فاضي اليوم؟
	Generated Response:	I'm very busy. I went I have an office انا مشغول كثير. رحت عندي مكتب
	Message:	What are you planting in your garden ? إيش قاعد تزرع في حديقتك؟
	Generated Response:	I'm busy studying. I'm learning some academic subjects انا مشغول في الدراسة. اتعلم بعض المواد الاكاديمية

Table A.2: Lemon-picked examples generated using top-k sampling with $k = 50$.

BIBLIOGRAPHY

[1] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 3, pp. 1–32, 2020.

[2] S. Roller, Y.-L. Boureau, J. Weston, A. Bordes, E. Dinan, A. Fan, D. Gunning, D. Ju, M. Li, S. Poff, *et al.*, "Open-domain conversational agents: Current progress, open problems, and future directions," *arXiv preprint arXiv:2006.12442*, 2020.

[3] M. Czerwinski, J. Hernandez, and D. McDuff, "Building an AI that feels: AI systems with emotional intelligence could learn faster and be more helpful," *IEEE Spectrum*, vol. 58, no. 5, pp. 32–38, 2021.

[4] Ö. N. Yalçın, "Empathy framework for embodied conversational agents," *Cognitive Systems Research*, vol. 59, pp. 123–132, 2020.

[5] Ö. N. Yalçın and S. DiPaola, "A computational model of empathy for interactive agents," *Biologically Inspired Cognitive Architectures*, vol. 26, pp. 20–25, 2018.

[6] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston, "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*, 2021, pp. 300–325.

[7] P. Zhong, C. Zhang, H. Wang, Y. Liu, and C. Miao, "Towards persona-based empathetic conversational models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6556–6566.

[8] Ö. N. Yalçın and S. DiPaola, "M-path: A conversational system for the empathic virtual agent," in *Biologically Inspired Cognitive Architectures Meeting*, Springer, 2019, pp. 597–607.

[9] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2018.

[10] 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.

[11] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung, "CAiRE: an end-to-end empathetic chatbot.," in *AAAI*, 2020, pp. 13 622–13 623.

[12] ——, "Caire: An end-to-end empathetic chatbot," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 622–13 623.

[13] V.-K. Tran and M. Le Nguyen, "Dual latent variable model for low-resource natural language generation in dialogue systems," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 2018, pp. 21–30.

[14] J. Gu, Y. Wang, Y. Chen, V. O. Li, and K. Cho, "Meta-learning for low-resource neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3622–3631.

[15] K. Kann, J. Bjerva, I. Augenstein, B. Plank, and A. Søgaard, "Character-level supervision for low-resource pos tagging," in *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, 2018, pp. 1–11.

[16] A. Otegi, A. Agirre, J. A. Campos, A. Soroa, and E. Agirre, "Conversational question answering in low resource scenarios: A dataset and case study for basque," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 436–442.

[17] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," *arXiv preprint arXiv:2010.12309*, 2020.

[18] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Marseille, France: European Language Resource Association, May 2020, pp. 9–15. [Online]. Available: https://www.aclweb.org/anthology/2020.osact-1.2.

[19] D. A. Ali and N. Habash, "Botta: An arabic dialect chatbot," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 2016, pp. 208–212.

[20] 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)*, IEEE, 2014, pp. 227–237.

[21] A. Fadhil and A. AbuRa'ed, "OlloBot - 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)*, Sep. 2019, pp. 295–303.

[22] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[23] J. Lu, P. Gong, J. Ye, and C. Zhang, "Learning from very few samples: A survey," *arXiv preprint arXiv:2009.02653*, 2020.

[24] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[25] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 2, pp. 151–166, 2020.

[26] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[27] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[28] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[29] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[30] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," *arXiv preprint arXiv:1803.00676*, 2018.

[31] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[32] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 221–230.

[33] E. Bart and S. Ullman, "Cross-generalization: Learning novel classes from a single example by feature replacement," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 672–679.

[34] Y.-X. Wang and M. Hebert, "Learning from small sample sets by combining unsupervised meta-training with cnns," *Advances in Neural Information Processing Systems*, vol. 29, pp. 244–252, 2016.

[35] ——, "Learning to learn: Model regression networks for easy small sample learning," in *European Conference on Computer Vision*, Springer, 2016, pp. 616–634.

[36] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[38] C. Pan, J. Huang, J. Gong, and X. Yuan, "Few-shot transfer learning for text classification with lightweight word embedding based models," *IEEE Access*, vol. 7, pp. 53 296–53 304, 2019.

[39] P. Fivez, S. Suster, and W. Daelemans, "Scalable few-shot learning of robust biomedical name representations," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 23–29.

[40] H. Mulki, H. Haddad, M. Gridach, and I. Babaoğlu, "Syntax-ignorant n-gram embeddings for dialectal arabic sentiment analysis," *Natural Language Engineering*, pp. 1–24,

[41] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 440–450.

[42] A. Rios and R. Kavuluru, "Few-shot and zero-shot multi-label learning for structured label spaces," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3132–3142.

[43] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-shot charge prediction with discriminative legal attributes," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 487–498.

[44] S. Paul, P. Goyal, and S. Ghosh, "Automatic charge identification from facts: A few sentence-level charge annotations is all you need," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1011–1022.

[45] M. Falis, M. Pajak, A. Lisowska, P. Schrempf, L. Deckers, S. Mikhael, S. Tsaftaris, and A. O'Neil, "Ontological attention ensembles for capturing semantic concepts in icd code prediction from clinical text," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 2019, pp. 168–177.

[46] D. Song, A. Vold, K. Madan, and F. Schilder, "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training," *Information Systems*, p. 101 718, 2021.

[47] K. He, Y. Yan, H. Xu, S. Liu, Z. Liu, and W. Xu, "Learning label-relational output structure for adaptive sequence labeling," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.

[48] J. Li, T. Tang, W. X. Zhao, and J.-R. Wen, "Pretrained language models for text generation: A survey," *arXiv preprint arXiv:2105.10311*, 2021.

[49] Z. Chen, H. Eavani, W. Chen, Y. Liu, and W. Y. Wang, "Few-shot NLG with pre-trained language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 183–190.

[50] H. Gong, Y. Sun, X. Feng, B. Qin, W. Bi, X. Liu, and T. Liu, "TableGPT: Few-shot table-to-text generation with table structure reconstruction and content matching," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1978–1988.

[51] J. Li, T. Tang, W. X. Zhao, Z. Wei, N. J. Yuan, and J.-R. Wen, "Few-shot knowledge graph-to-text generation with pretrained language models," *arXiv preprint arXiv:2106.01623*, 2021.

[52] Z. Yang, J. Yang, C. Xu, *et al.*, "Low-resource response generation with template prior," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 1886–1897.

[53] B. Peng, C. Zhu, C. Li, X. Li, J. Li, M. Zeng, and J. Gao, "Few-shot natural language generation for task-oriented dialog," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 172–182.

[54] I. Shalyminov, S. Lee, A. Eshghi, and O. Lemon, "Few-shot dialogue generation without annotated data: A transfer learning approach," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 2019, pp. 32–39.

[55] C. Xia, C. Zhang, H. Nguyen, J. Zhang, and P. Yu, "CG-BERT: Conditional text generation with BERT for generalized few-shot intent detection," *arXiv preprint arXiv:2004.01881*, 2020.

[56] S. Coope, T. Farghly, D. Gerz, I. Vulić, and M. Henderson, "Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 107–121.

[57] C. Xia, C. Xiong, S. Y. Philip, and R. Socher, "Composed variational natural language generation for few-shot intents," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 3379–3388.

[58] J. Gu, Y. Wang, Y. Chen, V. O. Li, and K. Cho, "Meta-learning for low-resource neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3622–3631.

[59] S. Dingliwal, S. Gao, S. Agarwal, C.-W. Lin, T. Chung, and D. Hakkani-Tur, "Few shot dialogue state tracking using meta-learning," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 1730–1739.

[60] M. Yu, X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauro, H. Wang, and B. Zhou, "Diverse few-shot text classification with multiple metrics," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1206–1215.

[61] Y. SONG, Z. Liu, W. Bi, R. Yan, and M. Zhang, "Learning to customize model structures for few-shot dialogue generation tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5832–5841.

[62] S. Deng, N. Zhang, J. Kang, Y. Zhang, W. Zhang, and H. Chen, "Meta-learning with dynamic-memory-based prototypical network for few-shot event detection," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 151–159.

[63] J. Li, B. Chiu, S. Feng, and H. Wang, "Few-shot named entity recognition via meta-learning," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[64] W. Yin, "Meta-learning for few-shot natural language processing: A survey," *arXiv preprint arXiv:2007.09604*, 2020.

[65] T. Bansal, R. Jha, T. Munkhdalai, and A. McCallum, "Self-supervised meta-learning for few-shot natural language classification tasks," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 522–534.

[66] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 928–939.

[67] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[68] H.-Y. Shum, X.-d. He, and D. Li, "From Eliza to XiaoIce: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10–26, 2018.

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

[70] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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*, 2020, pp. 7871–7880.

[71] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "Mt5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 483–498.

[72] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and W. B. Dolan, "DIALOGPT: Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 270–278.

[73] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "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.

[74] 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?" In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2204–2213.

[75] N. Asghar, I. Kobyzev, J. Hoey, P. Poupart, and M. B. Sheikh, "Generating emotionally aligned responses in dialogues using affect control theory," *arXiv preprint arXiv:2003.03645*, 2020.

[76] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of XiaoIce, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[77] S. AlHumoud, A. Al Wazrah, and W. Aldamegh, "Arabic chatbots: A survey," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 535–541, 2018.

[78] H. Mozannar, K. E. Hajal, E. Maamary, and H. Hajj, "Neural arabic question answering," *arXiv preprint arXiv:1906.05394*, 2019.

[79] O. ElJundi, W. Antoun, N. El Droubi, H. Hajj, W. El-Hajj, and K. Shaban, "HULMonA: The universal language model in Arabic," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 68–77. DOI: 10.18653/v1/W19-4608. [Online]. Available: https://www.aclweb.org/anthology/W19-4608.

[80] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for arabic," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, 2016, pp. 11–16.

[81] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[82] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*, 2017, pp. 67–72.

[83] 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.

[84] T. Naous, C. Hokayem, and H. Hajj, "Empathy-driven arabic conversational chatbot," in *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, 2020, pp. 58–68.

[85] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[86] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[87] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[88] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*, PMLR, 2020, pp. 11 328–11 339.

[89] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 889–898.

[90] D. Ippolito, R. Kriz, J. Sedoc, M. Kustikova, and C. Callison-Burch, "Comparison of diverse decoding methods from conditional language models," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3752–3762.

[91] W. G. Parrott, *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[92] Ö. N. Yalçın, "Evaluating empathy in artificial agents," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019, pp. 1–7.

[93] M. Abdul-Mageed, C. Zhang, A. Elmadany, and L. Ungar, "Beyond geolocation: Micro-dialect identification in diaglossic and code-switched environments," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5855–5876.

[94] M. Abdul-Mageed, C. Zhang, H. Bouamor, and N. Habash, "Nadi 2020: The first nuanced arabic dialect identification shared task," in *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, 2020, pp. 97–110.

[95] H. Bouamor, S. Hassan, and N. Habash, "The madar shared task on arabic fine-grained dialect identification," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 199–207.

[96] R. Eskander, N. Habash, O. Rambow, and N. Tomeh, "Processing spontaneous orthography," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 585–595.

[97] A. Fadhil *et al.*, "OlloBot-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)*, 2019, pp. 295–303.

[98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[99] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for arabic," *arXiv preprint arXiv:2101.01785*, 2020.

[100] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 986–995.

[101] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[102] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[103] D. Ippolito, R. Kriz, J. Sedoc, M. Kustikova, C. Callison-Burch, R. Kriz, E. Miltsakaki, M. Apidianaki, C. Callison-Burch, J. Hewitt, *et al.*, "Comparison of diverse decoding methods from conditional language models," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, Association for Computational Linguistics, 2018.

[104] S. Basu, G. S. Ramachandran, N. S. Keskar, and L. R. Varshney, "Mirostat: A neural text decoding algorithm that directly controls perplexity," in *International Conference on Learning Representations*, 2020.

[105] M. Djandji, F. Baly, W. Antoun, and H. Hajj, "Multi-task learning using AraBert for offensive language detection," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Marseille, France: European Language Resource Association, May 2020.

[106] H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, Eds., *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Marseille, France: European Language Resource Association, May 2020.

[107] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, *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*, 2021, pp. 300–325.

[108] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 3, pp. 1–32, 2020.

[109] D. Ippolito, R. Kriz, J. Sedoc, M. Kustikova, and C. Callison-Burch, "Comparison of diverse decoding methods from conditional language models," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3752–3762.

[110] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A survey on retrieval-augmented text generation," *arXiv preprint arXiv:2202.01110*, 2022.

[111] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[112] Q. Zhu, L. Cui, W. Zhang, F. Wei, and T. Liu, "Retrieval-enhanced adversarial training for neural response generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3763–3773.

[113] Y. Wu, F. Wei, S. Huang, Y. Wang, Z. Li, and M. Zhou, "Response generation by context-aware prototype editing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7281–7288.

[114] D. Cai, Y. Wang, W. Bi, Z. Tu, X. Liu, W. Lam, and S. Shi, "Skeleton-to-response: Dialogue generation guided by retrieval memory," 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)*, 2019, pp. 1219–1228.

[115] R. Yan, Y. Song, and H. Wu, "Learning to respond with deep neural networks for retrieval-based human-computer conversation system," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 55–64.

[116] G. Pandey, D. Contractor, V. Kumar, and S. Joshi, "Exemplar encoder-decoder for neural conversation generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1329–1338.

[117] H. Peng, A. Parikh, M. Faruqui, B. Dhingra, and D. Das, "Text generation with exemplar-based adaptive decoding," 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)*, 2019, pp. 2555–2565.

[118] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact aware neural abstractive summarization," in *thirty-second AAAI conference on artificial intelligence*, 2018.

[119] J. Gu, Y. Wang, K. Cho, and V. O. Li, "Search engine guided neural machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[120] K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang, "Generating sentences by editing prototypes," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 437–450, 2018.

[121] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781.

[122] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[123] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 857–16 867, 2020.