Empathy-driven Arabic Conversational Chatbot

Tarek Naous, Christian Hokayem, and Hazem Hajj
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
{tnn11,cph04,hh63}@aub.edu.lb

Abstract

Conversational models have witnessed a significant research interest in the last few years with the advancements in sequence generation models. A challenging aspect in developing human-like conversational models is enabling the sense of empathy in bots, making them infer emotions from the person they are interacting with. By learning to develop empathy, chatbot models are able to provide human-like, empathetic responses, thus making the human-machine interaction close to human-human interaction. Recent advances in English use complex encoder-decoder language models that require large amounts of empathetic conversational data. However, research has not produced empathetic bots for Arabic. Furthermore, there is a lack of Arabic conversational data labeled with empathy. To address these challenges, we create an Arabic conversational dataset that comprises empathetic responses. However, the dataset is not large enough to develop very complex encoder-decoder models. To address the limitation of data scale, we propose a special encoder-decoder composed of a Long Short-Term Memory (LSTM) Sequence-to-Sequence (Seq2Seq) with Attention. The experiments showed success of our proposed empathy-driven Arabic chatbot in generating empathetic responses with a perplexity of 38.6, an empathy score of 3.7, and a fluency score of 3.92.

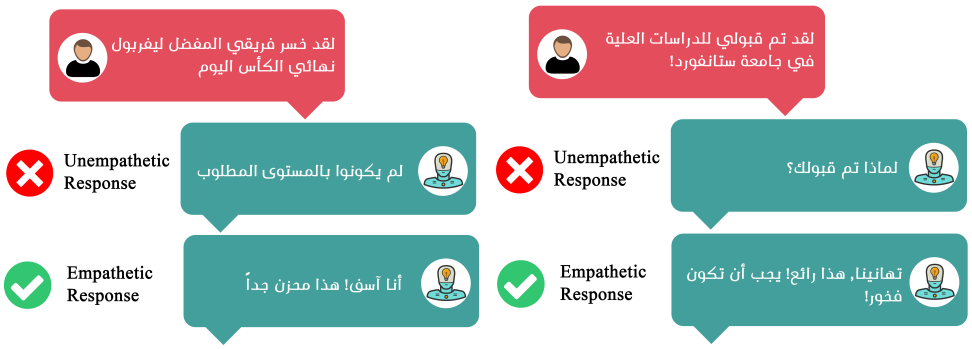
1 Introduction

Empathy is described as the ability of recognizing others' state of mind and making sense of their feelings. Empathetic behavior is provoked after being exposed to others' emotional states (Yalçın, 2020). Empathy is an innate capacity in most human beings, and is also described as a responsive and spontaneous act of copying of an implied feeling. Empathy in humans triggers a sense of concern for others, leading to appropriate emotional reactions that impose a positive effect on interacting individuals. For instance, empathetic behavior is applicable to situations such as acknowledging others' pain, showing interest, gratitude, being supportive, or providing encouragement.

Building chatbots that can exhibit empathetic behavior becomes a necessary step towards building emotionally intelligent conversational systems (Yalçın and DiPaola, 2018). Such a trait allows conversational systems to be perceived as genuine and warm by users, rather than oblivious and boorish. This characteristic is highly desirable since it could boost user satisfaction in various chatbot applications. Hence, the objective of this work is to develop an empathetic chatbot for the Arabic language. An example of the desired chatbot with empathetic behavior is illustrated in Fig. 1 that shows the difference between empathetic and unempathetic responses. For instance, in Fig 1a the chatbot infers a feeling of sadness in the user's text and provides an empathetic response to comfort the user, while in Fig 1b the chatbot infers a feeling of pride in the user's statement and thus congratulates them.

Recent advances in Artificial Intelligence (AI) and Natural Language Processing (NLP) have made the development of such systems possible. Specifically, the introduction of Sequence-to-Sequence (Seq2Seq) models and Transformer networks have improved performance significantly. Empathetic chatbots have grabbed interest in recent years with many sequence architectures proposed to generate

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 58–68
Barcelona, Spain (Online), December 12, 2020



(a) User feeling sad (b) User feeling proud

Figure 1: Examples for Empathetic Responding

empathetic responses given a user's emotional statement (Zhong et al., 2020; Zhou et al., 2020). The recent attempts in building empathetic chatbots for English have focused on training sequence generation models on a dataset of empathetic conversations (Rashkin et al., 2018), where models are trained on input/output sequences of speaker statements and their corresponding empathetic responses. The dataset used for English is based on empathetic conversations between a speaker and a listener. The speaker describes a situation he previously experienced, while the listener infers the emotional state of the speaker and provides a suitable empathetic response. Hence, by training the model on such samples of speaker statements and listener empathetic responses, the chatbot learns to develop a sense of empathy and provides suitable responses to any given context.

Despite the various works presented in the literature on empathetic chatbots for English (Shin et al., 2019; Ghandeharioun et al., 2019; Asghar et al., 2020), no work has previously addressed the problem of building such a model for the Arabic language. This is mainly due to the difficulties that arise when dealing with a morphologically rich language such as Arabic. Another important reason is the scarcity of resources available for Arabic compared with the English language, including corpora, tools, and pre-trained models. To address these challenges, we propose a neural sequence generation model based on Seq2Seq with LSTM units combined with Attention. We select this model over state-of-the-art transformer models due to their low computational cost and suitability for smaller datasets that are more feasible for Arabic. Additionally, we create a corpus for empathetic conversations in Arabic by translating datasets available for English. The developed model successfully exhibited empathetic behavior and provided emotional responses to the input of users in Arabic.

The rest of this paper is organized as follows: Section 2 summarizes the recent literature on empathetic chatbots. Section 3 describes the Arabic dataset created for empathetic chatbots and presents the proposed Seq2Seq model. Experiments and results are presented in Section 4. Concluding remarks and future directions follow in Section 5.

2 Related Work

In this section, we start by reviewing the recent literature on empathetic chatbots for the English language, including datasets and models used. We then review the current state of Arabic chatbots and highlight the existing gaps.

2.1 Empathetic English Chatbots

English empathetic chatbots have been of interest over the last few years. Recently, the first dataset for empathetic conversations dubbed EmpatheticDialogues was introduced by Rashkin et al. (2018). 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, Lin et

al. (2020) proposed an improved model which employed a Generative Pre-trained Transformer (GPT). This model was pre-trained on the BooksCorpus dataset (Zhu et al., 2015) 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 (Zhang et al., 2018) 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, Shin et al. (2020) 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. Asghar et al. (2020) 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. Zhou et al. (2020) 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.

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 (Hijjawi et al., 2014). Another work is BOTTA, a retrieval-based model supporting specifically the Egyptian dialect (Ali and Habash, 2016). For the medical domain, Ollobot is another rule-based chatbot which presents health tracking and support (Fadhil and AbuRa'ed, 2019). Overall, in a survey conducted by AlHumoud et al. (2018), 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 developed by Mozannar et al. (2019), another one is the Arabic language model developed by Antoun et al. (2020). The success of these works, as well as the work of ElJundi et al. (2019) 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 Proposed Method

In this section, we present the proposed model for Arabic empathy-driven conversational bots and the dataset we created. We start by describing the details of the proposed encoder-decoder model in Subsection 3.1. We then present the dataset created for Arabic empathetic chatbots in Subsection 3.2, which we used to train our proposed model.

3.1 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. 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.

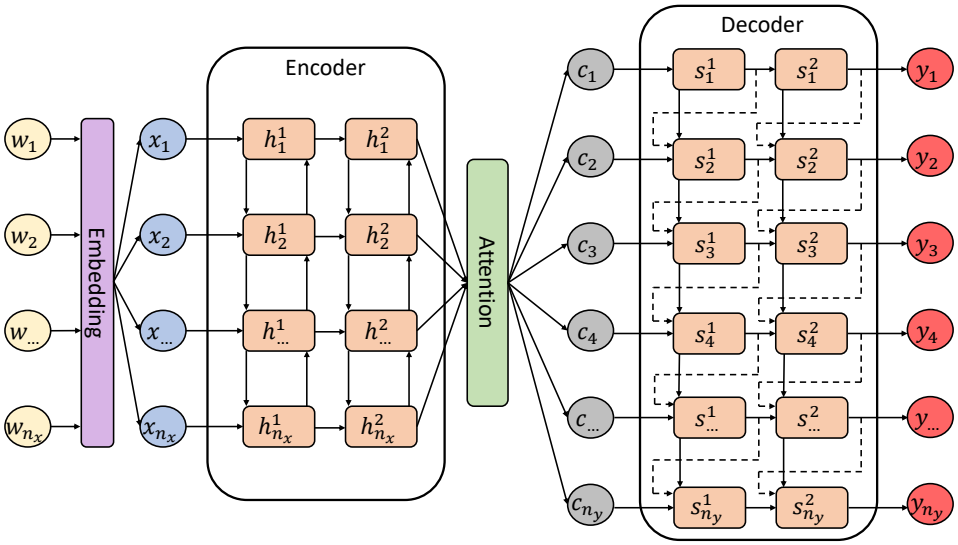


Figure 2: Architecture of the proposed Seq2Seq model with Attention

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 (Abdelali et al., 2016) 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 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 (Bahdanau et al., 2014), 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.

3.1.1 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:

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

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:

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

3.1.2 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:

$$\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_k^{n_x} \exp(e_{ki})} \tag{3}$$

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:

$$e_{ji} = f_{NN}(s_{j-1}, h_i) \tag{4}$$

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:

$$c_j = \sum_{j=1}^{n_x} \alpha_{ji} h_i \tag{5}$$

3.1.3 Decoder

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

$$\begin{aligned} s_j^1 &= \text{LSTM}(c_j, s_{j-1}^1, s_{j-1}^2) \\ s_j^2 &= \text{LSTM}(s_j^1, s_{j-1}^2, y_{j-1}) \end{aligned} \tag{6}$$

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.

$$p(\mathbf{y}) = \prod_{j=1}^{n_y} p(y_j / y_1, y_2, \ldots, y_{j-1}, c) \tag{7}$$

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

3.2 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 (Rashkin et al., 2018), 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

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

English to Arabic. A sample conversation is provided in Table 1, showing the original English sentences and their Arabic translations.

Hello مرحبا
Hello! I am in such a good mood since I got my new home مرحبا! أنا في مزاج جيد منذ أن حصلت على منزلي الجديد
Funny - we just built a house where we used to go camping when I was a kid مضحك ـ قمنا ببناء منزل حيث كنّا نذهب للتخييم عندما كنت طفلا
I have a ton of backyard space that we have plans for. Camping would be a fun one لديّ الكثير من مساحة الفناء الخلفي اللتي نخطط لها. التّخييم سيكون ممتعا
That's wonderful! Do you have any plans to have a fire? هذا رائع. هل لديك خطط لإشعال النّار؟
I want to look in to getting a nice fire pit for the house أريد أن أنظر للحصول على حفرة نار جميلة

Table 1: Sample conversation from the created dataset.

To evaluate the quality of the dataset[2], 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 of training the proposed empathetic conversational model. A few unreasonable samples are shown in Table 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 2: Examples of unreasonable translations.

4 Experiments and Results

In this section, we start by defining the experimental setup in Subsection 4.1, including how the dataset is split and what model configurations is trained. We analyze the results in Subsection 4.2. We then present the results of the human evaluations in Subsection 4.3 and provide a discussion on the strengths and shortcoming of the proposed model.

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

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, presented earlier in Section 3 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 (Klein et al., 2017) toolkit which is commonly used for neural sequence learning.

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 show the variation of the validation PPL over 8000 training steps, for the three choices of d. As observed in Fig. 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, 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.

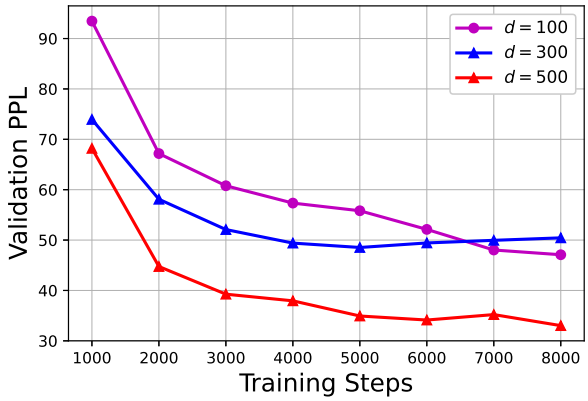


Figure 3: Validation PPL curves for several word embedding dimensions d

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

Table 3: Performance of the models on the test set in terms of PPL and BLEU score.

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 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.

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

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

Table 5: Sample generated responses by the proposed model. Generated responses are shown using both beam search and random sampling decoding strategies at inference.

For further analysis of the model's performance, we show in Table 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 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 (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 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 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.

5 Conclusion

In this paper, we proposed the first model for Arabic empathetic conversational bots and a dataset of empathetic conversations in Arabic. Our proposed model is a Seq2Seq model with LSTM units combined with attention. The dataset created was translated from the EmpatheticDialogues dataset available in English and showed an accuracy of 96% on a sample of the data, which was deemed as sufficient for the purpose of training conversational bots. Upon experimenting with several model configurations, the proposed model with an embedding dimension of 500 reached state of the art performance for Arabic with a PPL of 38.6 and a BLEU score of 0.5. Human evaluation of the generated responses also validated the success of the proposed model, which reached an average Empathy score of 3.7 and an average Fluency score of 3.92. Our results were promising and showed the ability of the proposed model in inferring speaker emotions and generating empathetic responses. However, the model showed average score in Relevance indicating that the response may sometimes go off topic. The limitations of this model are mainly due to the small size of the dataset created. Therefore, our future directions include creating a large conversational dataset for the Arabic language. The larger dataset will also enable the exploration of complex sequence generation models such as Transformers and pre-trained models.

Acknowledgement

This work has been funded by the American University of Beirut (AUB) University Research Board (URB).

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. 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*, pages 11–16.

Sarah AlHumoud, Asma Al Wazrah, and Wafa Aldamegh. 2018. Arabic chatbots: A survey. *International Journal of Advanced Computer Science and Applications*, 9(8):535–541.

Dana Abu Ali and Nizar Habash. 2016. Botta: An arabic dialect chatbot. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 208–212.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. 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*, pages 9–15, Marseille, France, May. European Language Resource Association.

Nabiha Asghar, Ivan Kobyzev, Jesse Hoey, Pascal Poupart, and Muhammad Bilal Sheikh. 2020. Generating emotionally aligned responses in dialogues using affect control theory. *arXiv preprint arXiv:2003.03645*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Obeida ElJundi, Wissam Antoun, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2019. hULMonA: The universal language model in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 68–77, Florence, Italy, August. Association for Computational Linguistics.

Ahmed Fadhil and Ahmed AbuRa'ed. 2019. 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)*, pages 295–303, Varna, Bulgaria, September. INCOMA Ltd.

Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. Emma: An emotion-aware wellbeing chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE.

Mohammad Hijjawi, Zuhair Bandar, Keeley Crockett, and David Mclean. 2014. ArabChat: an arabic conversational agent. In *2014 6th International Conference on Computer Science and Information Technology (CSIT)*, pages 227–237. IEEE.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. CAiRE: an end-to-end empathetic chatbot. In *AAAI*, pages 13622–13623.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. *arXiv preprint arXiv:1906.05394*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *arXiv preprint arXiv:1906.08487*.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7989–7993. IEEE.

Özge Nilay Yalçın and Steve DiPaola. 2018. A computational model of empathy for interactive agents. *Biologically Inspired Cognitive Architectures*, 26:20–25.

Özge Nilay Yalçın. 2020. Empathy framework for embodied conversational agents. *Cognitive Systems Research*, 59:123–132.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Peixiang Zhong, Yan Zhu, Yong Liu, Chen Zhang, Hao Wang, Zaiqing Nie, and Chunyan Miao. 2020. Endowing empathetic conversational models with personas. *arXiv preprint arXiv:2004.12316*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. 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*, pages 19–27.