Neural Encoder-Decoder based Urdu Conversational Agent

Mehreen Alam
Department of Computer Science
NUCES
Islamabad, Pakistan
mehreen.alam@nu.edu.pk

Sibt ul Hussain
Department of Computer Science
NUCES
Islamabad, Pakistan
sibtul.hussain@nu.edu.pk

***Abstract*—Conversational agents have very much become part of our lives since the renaissance of neural network based ”neural conversational agents”. Previously used manually annotated and rule based methods lacked the scalability and generalization capabilities of the neural conversational agents. A neural conversational agent has two parts: at one end an encoder understands the question while the other end a decoder prepares and outputs the corresponding answer to the question asked. Both the parts are typically designed using recurrent neural network and its variants and trained in an end-to-end fashion. Although conversation agents for other languages have been developed, Urdu language has seen very less progress in building of conversational agents. Especially recent state of the art neural network based techniques have not been explored yet. In this paper, we design an attention driven deep encoder-decoder based neural conversational agent for Urdu language. Overall, we make following contributions we (i) create a dataset of 5000 question-answer pairs and (ii) present a new deep encoder-decoder based conversational agent for Urdu language. For our this work we limit the knowledge base of our agent to general knowledge regarding Pakistan. Our best model has the BLEU score of 58 and gives syntactically and semantically correct answers in majority of the cases.**

***Index Terms*—deep learning, machine learning, conversational model, sequence to sequence, chatbot, Urdu language**

I. Introduction

Advances in end to end training of deep neural networks have led to huge successes in a variety of fields including automatic speech recognition, natural language processing and computer vision. One such area that has seen a significant leap in performance is the area of conversational modeling where humans put up a conversation with a conversational agent just like they are talking with human fellows. For a neural agent, to be to able to do conversation with an human at the level of human is extremely difficult task due to inclusion of all the previous context, diverse range of questions, pro-activeness and emotions. Thus neural conversation agents are either constrained to the knowledge area to a specific domain or made to answer questions directly without keeping track of previous context of conversation. Even after the addition of these restrictions, there is still a lot of work needed to be done to achieve human level performance.

In this work our goal is to build on the recent advancements in the domain of conversational agents and build an Urdu conversational agent. Urdu is understood by at least 20 million people in the world and to date, to the best of our knowledge, there is no publicly available conversational model available. Neither there has been any publicly available dataset in Urdu that could be used for building of the conversational model. So here we have chosen the domain of general knowledge of Pakistan and limit the conversation agent context history to a single question-answer pair only. Precisely, we transform our question answering problem to the deep learning technique of sequence to sequence proposed by (1). Overall, we make the following contributions:

1) Create a parallel corpus of 5000 question-answers pairs in Urdu language in the domain of general knowledge of Pakistan, and
2) Design a model for the Urdu based conversational agent in the domain of general knowledge of Pakistan.

The model we have used for our system is based on bi-directional encoder and attention based decoder. Bi-directional encoders learn in the forward as well as the backward direction, thereby increasing the learning capacity of the encoder (14). Attention mechanisms show remarkable improvement in quality as they tend to focus on the relevant parts of the question while producing the answers (14). We have used long short-term memory networks rather than vanilla recurrent neural network or its variations, since LSTM overcome the problem of vanishing gradients and give better mapping even for longer questions and answers according to (3). For our problem, single layer of encoder and decoder gave the most optimal results qualitatively and quantitatively. The evaluation metric used for judging the quality of answers generated for every question is the standard Bilingual Evaluation Understudy (BLEU) proposed by (2). The aforementioned evaluation criteria is relevant to our model as we map questions asked as the input sequence while the answers are mapped onto the output sequence.

The layout of the paper is as follows. We present background study and motivation in the next section. Section III throws light on the model architecture in detail, followed by the how we built the dataset in section IV. We give the details of our experimental settings in section V and a thorough discussion on the results of our model in section VI. We finally conclude our paper in section VII and highlight the future directions.

II. Background Study and Motivation

Humans use currently available textual conversational agents, like A.L.I.C.E. and Miksuku to help them in completing specific tasks or merely to put up a conversation. Work in this domain picked up pace with the introduction of increased compute power and deeper networks for enhanced learning. Conventional techniques are usually template-based or heuristic based as have been used by (4; 5). Currently, data-driven end-to-end training using deep learning techniques is used by (6; 7; 8; 10).

Since the problem is too diverse and complex, work has been done by looking at only a specific portion of the problem. (11) categorizes the problem into task-oriented and non-task oriented conversational agents. The former works as an assistant to the user for specific trained tasks like flight reservations, hotel booking. Non-task oriented systems are primarily built to converse with humans in a general way, chatbots is a common example. Major approaches used for building non-task oriented systems are: a) generative models which generate responses to the questions posed, and b) retrieval-based models that refer to a knowledge base before producing the answer. Generative models have the edge of conversing more human-like though the content may lack meaning. Retrieval based systems are more accurate but are more blunt and thus less likely to put up conversations closer to humans. Other ways of looking at the conversational modeling problem is either source to target mapping problem (12) or domain specific modeling problem (13).

One of the modern ways of looking at the problem of conversational model is to map it onto a machine translation task (12). After the introduction of sequence to sequence models by (1), many attempts have been made to extend the work by using this model (6; 8; 12). Variations to the encoder-decoder model are explored, for example, bi-directional encoder (14), attention based decoder (14) and use of word embeddings (17; 18) to enrich the learning, to name a few. However, to date, no such attempt has been made using any of the state-of-the-art deep learning based technique to work in line with building a conversational model for Urdu language. Most prominent ones like (15; 16) are task-oriented and use conventional methods. This leaves a big research gap in the domain of conversational modeling while in the context of Urdu language no prior attempt has been made.

To the best of or knowledge, our work is the first effort in the direction of building an Urdu based conversational agent in the domain of general knowledge of Pakistan using deep learning based sequence to sequence networks.

III. Model and its Architecture

We have used sequence to sequence model with bi-directional encoder and attention mechanism. Bi-directional encoder has the enhanced capacity as it learns in the forward as well as backward time direction. The attention mechanism spreads the learnt embedding for a question to a series of annotation vectors which at every step of generating the answer brings the focus to the relevant words in the question. This approach overcomes the bottleneck of relying on just one context vector where the question was embedded.

Specifically, our model has two RNNs: encoder and decoder. Encoder embeds the question of n words, $x_1, x_2, ..., x_n$, into a series of context vectors, $c_1, c_2, ..., c_n$,, while the decoder uses this vector to predict the answer of m words, $y_1, y_2, ..., y_m$. This is done by using the conditional probability, $p(y|x)$, such that question statement, x, has the answer statement, y and is explained in 1.

$$\log p(y|x) = \sum_{j=1}^{m} \log p(y_j|y_{<j}, c_j) \tag{1}$$

such that

$$p(y_j|y_{<j}, c_j) = softmax(w(h_j)) \tag{2}$$

where probability of every word to be predicted, y_j, given the previous words predicted, $y_{<j}$ and the context vector, c_j, is equal to taking softmax over w, which is the transformation function that maps onto the vector of the size of vocabulary and h_j is the LSTM output in the decoder network which is computed as

$$h_j = concat[f(x_j, h_{j-1}); f(x_j, h_{j+1})] \tag{3}$$

These annotations are used to find context vector, c_j, for every output word, y_j, such that

$$c_j = \sum_{j=1}^{n} \alpha_{ji} h_i \tag{4}$$

where the weight, α_{ji}, for each annotation vector, h_i, is:

$$\alpha_{ji} = \exp(g_{ji}) / \sum_{k=1}^{n} \exp(g_{jk}) \tag{5}$$

The context vector, c_j, for ever output word is used by the decoder to focus on only the relevant words from the input question. Instead of a single generic context vector, a distributed context vector improves the answering capability. Specifically, each of these annotation vectors convey to the model the extent to which every word in the question is related to the current output answer work.

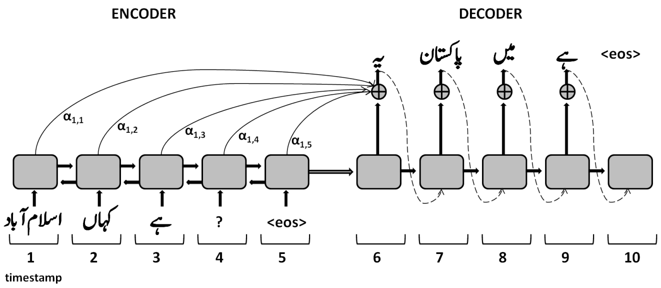

Figure 1: An example of how question-answer pair is mapped to bi-directional encoder and attention based decoder. We show context vector for the first word of the answer sequence.

g_{ji} is calculated from the annotation vector, h_i, and s_{j-1}, which is the output of the LSTM hidden unit just before emitting y_j.

IV. Dataset

As deep learning techniques are data-driven, good output quality heavily depends upon availability of a good dataset. Absence of any publicly available question answering data set, regardless of the domain, is the main reason for any progress done in this research direction. So, we had to build the dataset from scratch.

For us, it was a challenging task as there is no crowd sourcing facility available for Urdu language. The very famous Amazon Mechanical Turk is not operative in Pakistan. So, we came up with the novel idea of collecting data from 300 students of our university who entered questions into our database via a web portal we designed especially for this crowd sourcing facility. We gave them a list of topics in the context of Pakistan from which they could ask questions, for example about rivers, presidents, education, sports, provinces and economy, to name a few. This helped us capture the inherent diversity of the type and the way questions can be posed. This way we were able to get over 5000 general knowledge questions about Pakistan. Since the main part of a conversational agent is to understand the intent of the question which when found is mapped to one specific answer. To reach a wider audience, multiple questions and corresponding answer have been mentioned in english language, while originally everything was done completely in Urdu. For example, the answer for all the questions mentioned below is "Imran Khan":

1) Who is the Prime Minister of Pakistan?
2) Who is the Prime Minister?
3) Who is the current Prime Minister of Pakistan?
4) Prime Minister?
5) Prime Minister of Pakistan?

Data pre-processing step was intense as it needed manual as well as algorithmic techniques to remove the irrelevant questions and to fix the irregularities found in any question. It is worth mentioning that we did not fix the human errors as in real life humans are prone to enter questions such that may have grammatical or spelling errors but do make a sense overall. As an example, we list four versions of the question where each version has either a grammatical or a spelling mistake. All variations like these are part of our dataset. We have underlined the misspelled or incorrect words entered in the following examples:

- پاکستان کا وزیر اعظم کون حے؟
- پاکستان کا وزیر اعظم کیا ہے؟
- پاکستان کا وزیر اعظم کب ہے؟
- پاکستان کا وزیر اعضم کون ہے؟

This way we were able to capture the irregularities from a big majority of people which helped us make our model generic, scalable and focused in guessing the intent of the question while ignoring minor spelling or grammatical mistakes. After cleaning the data, we manually entered relevant answers to every question. Additionally, our model is independent of the sequence length which means it does not require the sequence lengths of the question and the corresponding answer be the same. This further makes our model more flexible and better prepared to understand the real-life diversities inherent in the language.

V. Experimental Settings

Thorough experimentation was done on GPU machines to find the most optimal parameters and hyper-parameters for the model. Cross validation was also done to make sure the model does not over-fit and gives us the most generic output. One layer deep LSTMs were chosen for both the encoder and the decoder after trying layer depths of 1,2,3,4 and 5. Hidden units in lstms, attention head and embedding dimension were all set to 128 after trying randomly between 64 to 2048 neurons all combinations in powers of 2. The sequence length was restricted to 30 with three buckets (0-10, 10-20, 20>). Adam was used the optimizer with learning rate of 0.0001 initially and 0.001 once the model had matured. Attention mechanism is used and encoders are used with bi-directional option. Our data was randomly shuffled and divided into train, validation and test set in the ratio of 3:1:1.

VI. Results and Discussions

A. *Quantitative Analysis*

After thorough experimentation and cross validation on the dataset created by us, our final model is the one presented earlier in Section III. It gave us the maximum BLEU score of 62 and minimum cross entropy loss of 0.5 on the validation dataset as seen from figure 2 and figure 3, outperforming any other combination of the network and its parameters within the time constraints. Thus, the model trained at 18K is the most optimal, we save that model and run it on the test set where it gives us the BLEU score of 58.

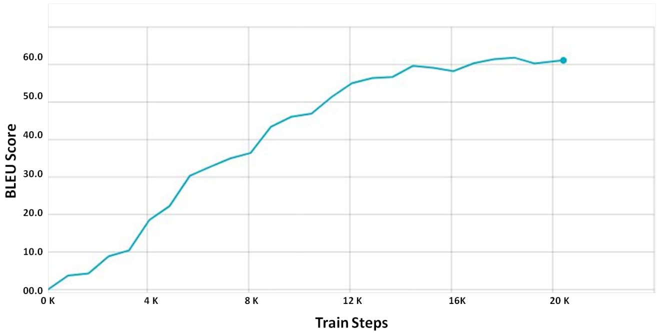

Figure 2: Bleu Score on Validation Dataset on a total of 25 K train steps. As can be seen that the model gives the top BLEU score of 62 at around 18K train steps.

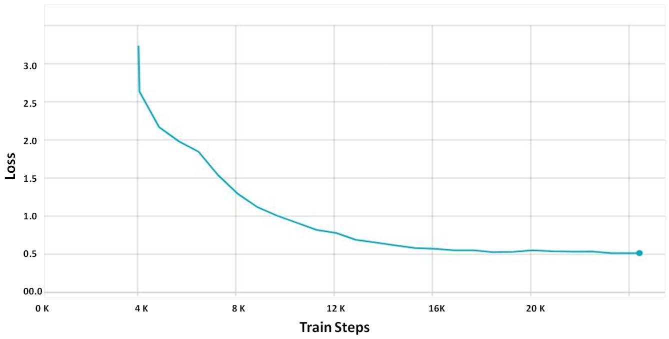

Figure 3: Loss on Validation Dataset on a total of 25 K train steps. In line with the BLEU score seen in figure above, the model has the lowest loss around 18K train steps.

B. *Qualitative Analysis*

A BELU score of 58 achieved by the model is a strong indication that the model is generic and generates accurate responses for majority of the cases. Table I shows the result. It can be observed from the examples in the table that the model digs the intent of the question even when question has an ambiguity. For example, the first question in table I which is asked differently than the regular way, our model has still grasped the intent of the question and answers exactly as a human shall do. Any question for which the model is not able to generate the relevant answer with a high probability, it refers to a one of the default answers requesting more clarity in posing the question as evident from question 7. Overall, it is very encouraging to observe that all answers are not only syntactically and grammatically correct but are also relevant and conceptually meaningful.

However, our model does makes mistakes on rare occasions. For analysis purposes, we have categorized the mistakes the model makes in two main ways: 1) conceptual, and 2) grammatical. Since deep learning techniques are data-driven, any question asked from a context the model has not seen before leads to an answer which may be conceptually irrelevant to the question asked but may be completely correct sentence otherwise, refer table II question no 3. Rest of the questions have been answered in conceptually correct manner with an error or two in appropriate preposition selection as highlighted by underlining the words.

Table I: Completely correct answers for the questions asked to our conversational agent

Answers	Questions	No.
اسلام آباد پاکستان کادارالحکومت ہے	اسلام آباد کیاہے؟	1
اسلام مذہب ہے	اسلام کیاہے؟	2
جیسمین پاکستان کا قومی پھول ہے	پاکستان کا قومی پھول؟	3
یہ پاکستان میں ہے	اسلام آباد کہاں ہے؟	4
میرے دوست ہو جو پاکستان کے بارے میں معلومات چاہتے ہو	میں کون ہوں؟	5
میں ایک بوٹ ہوں	تم کون ہو؟	6
کسی بھی الجھن سے بچنے کے لئے، براہ کرم مزید مخصوص رہیں	پاکستان کادارالحکومت؟	7

Table II: Answers having mistakes for the questions asked to our conversational agent

Answers	Questions	No.
قائداعظم پاکستان کا بانی ہے	پاکستان کا بانی کون ہے؟	1
آم پاکستان کے قومی پھل ہے	پاکستان کا قومی پھل کیاہے؟	2
یہ پاکستان کا قومی پھول ہے	کس نے ورلڈ کپ جیتا؟	3
ممنون حسین پاکستان کا صدر ہے	ممنون حسین کون ہے؟	4
پنجاب میں دارالحکومت ہے	پنجاب کہاں واقع ہے؟	5

VII. Conclusions

It was motivating to see that the first ever attempt to make an Urdu conversational agent in the domain of general knowledge of Pakistan has been so successful. We use modern state-of-the-art deep neural network based techniques to come up with the model for conversational agent that gives us solutions that are generic, scalable and able to give answers that are syntactically, conceptually and grammatically correct. It can safely be concluded that the model has developed sufficient cognitive abilities to respond with the precise contextual understanding. We have achieved the benchmark of 58 BLEU score. This score is the highest score attained by any conversational agent made in Urdu language and shall serve as a baseline for future work in that direction.

We plan to extend our work in many dimensions. We plan to work on creating a bigger dataset and also add more domains. We plan to explore the effects of using variants for RNNs, word embeddings; variants of encoders and use of different beam length for better answer quality.

References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks,"

Advances in neural information processing systems, pp. 3104-3112, 2014.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: a method for automatic evaluation of machine translation,” Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318, 2002.

[3] Sepp Hochreiter, and Jürgen Schmidhuber, “Long Short-Term Memory,” Neural Computation, pp. 1735-1780, 1997.

[4] L. Nio, S. Sakti, G. Neubig, T. Toda, M. Adriani, and S. Nakamura, “Developing non-goal dialog system based on examples of drama television,” Natural Interaction with Robots, Knowbots and Smartphones, Springer, New York, NY, 2014.

[5] D. Ameixa, L. Coheur, P. Fialho, and P. Quaresma, “Luke, I am our father: dealing with out-of-domain requests by using movies subtitles,” Intelligent Virtual Agents, Lecture Notes in Computer Science, vol 8637. Springer, Cham, 2014.

[6] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau, “Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models,” AAAI, pp. 3776–3784, 2016.

[7] L. Shang, Z. Lu, and H. Li, “Neural responding machine for short-text conversation,” In Proceedings of ACL, 2015.

[8] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Gao, B. Dolan, and J.-Y. Nie, “A neural network approach to context-sensitive generation of conversational responses,” In Proceedings of NAACL, 2015.

[9] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley, “A Knowledge-Grounded Neural Conversation Model,” arXiv, preprint arXiv:1702.01932, 2017.

[10] Jianfeng Gao, Michel Galley, and Lihong Li, “Neural Approaches to Conversational AI,” The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1371-1374, 2018.

[11] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang, “A survey on dialogue systems: recent advances and new frontiers,” ACM SIGKDD Explorations Newsletter, pp. 25–35, 2017.

[12] Alan Ritter, Colin Cherry, and William B. Dolan, “Data-driven response generation in social media,” Proceedings of the conference on empirical methods in natural language processing, pp. 583–593, 2011.

[13] Rafael E Banchs, and Haizhou Li, “IRIS: a chat-oriented dialogue system based on the vector space model,” Proceedings of the ACL 2012 System Demonstrations, pp. 37–42, 2012.

[14] D. Bahdanau, K. Cho, Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint in arXiv:1409.0473, 2014.

[15] Mohammed Kaleem, James O’Shea, and Keeley Crockett, “Development of umair the urdu conversational agent for customer service,” International Association of Engineers, 2014.

[16] Adnan A Arian, A Manzoor, K Brohi, K Haseeb, IA Halepoto, and IA Korejo, “Artificial intelligence mark-up language based written and spoken academic chatbots using natural language processing,” Sindh University Research Journal-SURJ (Science Series), pp. 153–158, 2018.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems, pp. 3111-3119, 2013.

[18] Jeffrey Pennington, Richard Socher, and Christopher Manning, “Glove: global vectors for word representation,” Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543, 2014.