Asian Journal of Research in Computer Science

1(3): 1-10, 2018; Article no.AJRCOS.42930

Conversation Modeling with Neural Network

Jivan Y. Patil[1*] and Girish P. Potdar[1]

[1] *Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India.*

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2018/v1i324738

Editor(s):

(1) Dr. Young Lee, Associate Professor, Department of Electrical Engineering and Computer Science, Frank H. Dotterweich College of Engineering, Texas A&M University-Kingsville, USA.

Reviewers:

(1) Ayush Dogra, Panjab University, India .

(2) Zlatin Zlatev, Trakia University, Bulgaria.

(3) Irshad Ullah, Pakistan.

Complete Peer review History: http://prh.sdiarticle3.com/review-history/25635

Received: 10th May 2018

Accepted: 14th July 2018

Published: 21st July 2018

Original Research Article

ABSTRACT

The ability to process, understand and interact in natural language carries high importance for building a Intelligent system, as it will greatly affect the way of communicating with the system. Deep Neural Networks (DNNs) have achieved excellent performance for many of machine learning problems and are widely accepted for applications in the field of computer vision and supervised learning. Although DNNs work well with availability of large labeled training set, it cannot be used to map complex structures like sentences end-to-end. Existing approaches for conversational modeling are domain specific and require handcrafted rules. This paper proposes a simple approach based on use of neural networks' recently proposed sequence to sequence framework. The proposed model generates reply by predicting sentence using chained probability for given sentence(s) in conversation. This model is trained end-to-end on large data set. Proposed approach uses Attention to focus text generation on intent of conversation as well as beam search to generate optimum output with some diversity.Primary findings show that model shows common sense reasoning on movie transcript data set.

**Corresponding author: E-mail: jivan.y.patil@gmail.com;*

Keywords: Natural language processing; deep learning; Chatbots; natural language understanding; artificial intelligence; neural networks; reinforcement learning.

1 INTRODUCTION

Natural language is preferred as medium for communication among humans, making Natural language Understanding an important ability for the intelligent systems to have. The ability to comprehend and communicate in natural language is an important milestone in the advancement of Artificial intelligence. It will enable users to interact with the system more naturally rather than having them go through option based User Interface which is rather very robotic. An intelligent system with ability to comprehend and generate context aware responses can provide solution to complex problem like customer care. Natural Language Processing and Deep Learning plays important role in building such system which show intelligence. Machines are far more efficient at repetitive tasks than humans as they can perform same task again and again producing same results every time, allowing use of machines for such tasks to get better accuracy, consistency and scaling.

In recent years, availability of hardware with high computational power and software that can take advantage of such computational power implementation of the computation intensive solutions have become possible. Allowing researchers to explore application of neural networks for more complex tasks like text generation. Deep Neural Networks (DNNs) are powerful models which have already achieved good performance for different machine learning tasks like speech recognition, image captioning, image classification, text classification, etc.

Advances in field of machine learning with deep learning and neural networks have led to remarkable progress towards solutions for many complex learning tasks such as speech recognition, computer vision, and language processing. Recent advances in deep learning shows that neural networks are much more capable than just classification and can be used for mapping complex structures such as sentences. An good example of this ability is the sequence-to-sequence framework [1] used for machine translation. This feature of sequence-to-sequence framework to map complex structures end to end allows us to tackle difficult tasks where domain knowledge is not available and/or its very difficult to model rules manually[2].

Sequence-To-Sequence [1] model shows that RNNs with memory cells like LSTM [3] or GRU [4] can successfully map source language sentence to target language and can be trained in end-to-end fashion achieving good results for tasks like machine translation, sentence memorization. This approach can be extended for purpose of conversation modeling [2], by making task of conversation modeling into task of translating questions into answers. This baseline approach can successfully map dialogues and responses for simple tasks bus fails at complex tasks and lengthy sequences. Sequence-To-Sequence model can be further extended by using neural attention [5, 6, 7, 8] to keep text generation focused on the intent of the conversation as well as using beam search [9, 10] to select one of many possible sequences allowing model to be more diverse and human like at text generation task.

This approach uses vocabulary and word-embedding provided by word-to-vector model. But use of this dictionary approach limits the learning capacity of model to max word size defined by vocabulary size hyper parameter. This can be overcome with use of character level generative model instead of word level by using CNN softmax or character LSTM in combination with word level LSTM.

2 LITERATURE REVIEW

Artificial Intelligence (AI) have been a hot topic for past years and lot of research has been done in fields like machine translation, computer vision, pattern recognition,NLP, etc. Some of the work related to NLP, Deep Learning and Neural

Networks is noted in this section.

RNN-enco-deco-SMT[4] proposed new kind of neural network based on encoder and decoder for machine translation and new type of cell called Gated Recurrent Unit (GRU). The neural network proposed in article encodes source language sentence in fix length vector using encoder and decoder decodes this fix-length vector into variable length target language output sequence. The model proposed leads to improved BLEU score for statistical machine translation. attention-with-intention[5] proposed improvement over traditional encoder-decoder RNN models for conversation modeling by adding an attention network. Proposed model consists three RNNs, encoder encodes the input sentences and decoder that generates responses then there is newly added attention network that models intention of the conversation over time.

seq-2-seq[1] illustrated ability of multi-layer LSTM RNN to achieve good performance on Machine Translation tasks. Article also shows that reversing the input sequences yields in better representations of the dependencies in input sequence and expected output sequence as well as better mapping of source symbols to target symbols. Proposed approach produces good BLEU scores by itself and state-of-art results when coupled with other baseline systems. end-to-end[11] proposes a recurrent memory-based model with multi-hops and trains the same with standard gradient descent. Author then evaluates the model for question-answer task. Model attends to sequences in timely fashion by considering next relevant piece of information at each time step. Though the model outperforms the baseline unsupervised approaches, it is far inferior than supervised approach.

NMT-joint-align-translate[12] proposed a novel ”attention” mechanism for improvements in standard sequence-to-sequence models. Since not all information can be encoded in a single vector, author proposes an approach to overcomes this by introducing an attention vector based on weighted sum of the input hidden states. Then the attention weights are learned along with rest of the weights and biases in the network. Approach proposed in article enables model to focus on more important part of input sequence to generate output sequence. attention-MT[8] evaluates effect of various attention mechanisms for task of Machine Translation. The author proposes ”global” and ”local” attention models where attending over all source words and subset of source words respectively. batch-norm[13] proposed a technique to normalize unit activation and unit variance within network. Author shows that Batch Normalization leads to faster traiing and better accuracy for convolutional networks. It also reduces the need of dropout.

attentn-sent-summary[7] Extends sequence-to-sequence model for task of abstracting sentences summarization. Neural attention is added for soft alignment. neural-conv-model[2] applies sequence-to-sequence model for modeling conversations instead of Machine translation like in base paper. The proposed approach exploits the ability of RNNs to map compelx structures for purpose of modeling conversations. Author then trains model on IT-Helpdesk and Open Subtitles dataset.

show-attend-tell[6] tries to improve image captioning by allowing decoder to focus on specific part of image than entire image and finds correspondence between words and image patches. The RNN uses underlying CNN outputs as input to map objects in image patch with captions in knowledge base. char-CNN-txt-classify[14] evaluated deep Convolutional Neural Network (CNN) on large-scale text classification using one-hot encoding to achieve competitive performance.

char-decode-no-explicit-segmentn[15] illustrates LSTM model, the model unlike baseline LSTM model models makes use of per word character-level CNN outputs and highway layer. Since word embedding are completely avoided the resulting model has significantly fewer parameters while achieving better performance. char-aware-LM[**?**] evaluates use of character-level decoder in Natural Machine Translation, also proposes a bi-scale architecture with slow and fast layers in decoder. Both biscale and base character-level decoder models perform better than word-level models at Machine Translation. char-MT-no-segmentn[17] proposed a character-level Neural

Machine Translation model. Unlike other RNN based encoder decoder models for NMT, this model uses CNN with max-pooling for encoder while using highway layer to reduce size of source representation. Standard RNN is used as decoder.

CLSTM[18] proposes a Contextual LSTM (CLSTM) model which makes use of both the input word and context vector to predict next word. This model performs better at selecting next sentence and predicting next topic than the baseline models for same tasks. hierarchy-multi-RNN[19] proposed a new hierarchical RNN which learns both temporal and hierarchical representations without prior knowledge of structure or timescale of hierarchy. To achieve this binary boundary detectors are used at each layer which control propagation of information between neighboring layers.

3 MODEL

Figure illustrates high level diagram of the model in proposed work. It uses word embedding to represent relations between words and this is used to compute possible candidate word while generating output.

Proposed approach makes use of sequence-to-sequence framework [1]. The model is based on Recurrent Neural Network(RNN) which reads input at one token at a time while generating output at one token at time. To speedup training and obtaining better accuracy the true output sequence is fed to decoder for training and learning happens by back propagation. The model is trained to maximize cross-entropy of correct sentence. During inference greedy approach is used, where instead of true output sequence the output generated in previous step is fed to decoder as next token. Less greedy approach in form of Beam-Search is used to provide better output by taking into consideration multiple output paths instead of going with local maximum at each step like in greedy approach of vanilla decoder.

The Seq2Seq framework relies on the encoder-decoder paradigm. The encoder encodes the input sequence, while the decoder produces the target sequence. For example, consider dialogue pair is ”ABC”, ”WXYZ”. Then neural network can be used to map ”ABC” to ”WXYZ” as shown in figure

3.1 Encoder

Each word from input sequence is associated to a vector $w \in R^d$. Then, this sequence of vectors are processed by LSTM layer and last hidden state of LSTM is passed to the decoder network for further processing. This last hidden state outputs are known as final state and provide encoder representation of input vector.

3.2 Decoder

The encoded vector representation of input sequence **e** is used by the decoder network to generate output sequence word by word. Last hidden state of encoder e along with start of sentence indicator are given as input to decoder. Decoder LSTM computes next hidden state $h_0 \in R^h$. Then a function g is applied so that $s_0 := g(h_0) \in R^V$ is vector of same size as vocabulary. Then, apply Softmax to normalize it into vector of probabilities $p_0 \in R^V$. Each entry in p measures likelihood of each word in vocabulary being output token.

$$h_t = LSTM(h_{t-1}, w_{Ih}) \tag{3.1}$$

$$s_t = g(h_t) \tag{3.2}$$

$$p_t = Softmax(s_t) \tag{3.3}$$

$$i_t = argmax(p_t) \tag{3.4}$$

as proved in [1]. Decoding stops when end of statement token is generated.

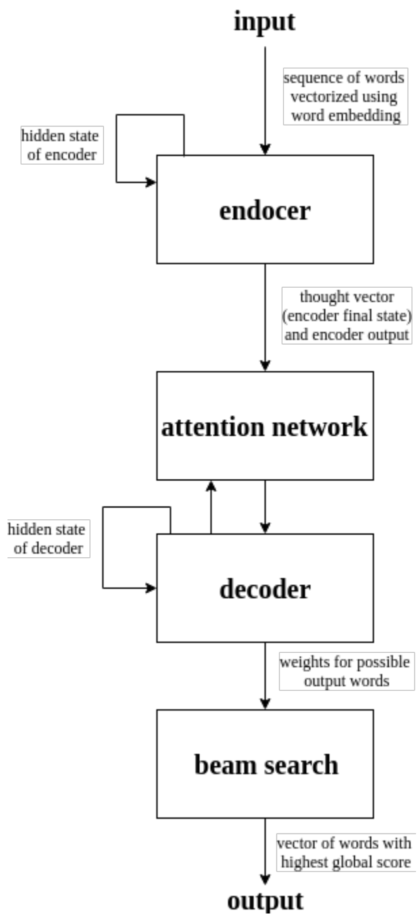
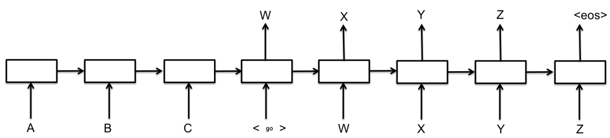


Fig. 1. High level diagram of generator model



[h]

Fig. 2. Example, use of seq2seq framework for conversational modeling

3.3 Decoder with Beam-Search

Greedy decoder suffers from local maxima problem, hence, giving less accurate answers. There is a better way of performing decoding, called Beam Search. Instead of only predicting the token with the best score, we keep track of k hypotheses (k is beam width). At each new time step we have V new possible tokens, resulting in k*V new hypothesis. We keep k best ones and repeat.

4 EXPERIMENTS

4.1 Data Set

The experiments are performed on cornell movie transcript dataset [20] and amazon product data [21]. The dataset contains 220,579 conversational exchanges between 10,292 pairs of movie characters. The dataset contains about 57K words, including boundary markers. All sentences are shuffled, duplicates removed, 40k words with highest frequency are chosen while all remaining are replaced with <UNK>token indicating that the word is out of vocabulary.

4.2 Model Setup

Perplexity is used as measure for analyzing model performance, which is the average per-word log-probability on holdout dataset: $e^{-\frac{1}{N}\sum_i \ln p_{w_i}}$. We compute perplexity by summing over all the words including the end of sentence token.

Cornell movie transcript dataset with vocabulary size of 40000 without any pre-processing and cleaning. The sentences are shuffled and are given as input to mdoel, start and end of statement is indicated with <SOS>and <EOS>tokens. For training maximum word length is set to 50.

4.3 Training Procedure

Using the predicted token as input to then next step during training increases errors as errors would accumulate over time-steps taken to generate output. This makes training slow if not impossible. To speed-up training as well as increase accuracy of model trained the actual output sequence is fed to the decoder LSTM while training while using the generated token for next step of decoding for inference.

The model is trained till convergence with ADAM optimizer using learning rate of 0.001 with learning rate decay of 0.999 and decay step size of 1000. Batch size of 32 is used while using 2 bidirectional LSTM layers with residual connections while using network size of 512. For decoding beam width is set to 30.

4.4 Model Evaluation

Perplexity : it is the average per-word log-probability on holdout dataset. Perplexity is computed by summing over all the words including the end of sentence token. [22]:

$$e^{-\frac{1}{N}\sum_i \ln p_{w_i}} \tag{4.1}$$

Training loss : Cross-Entropy loss [23]

$$H(p,q) = -\sum_x p(x)\log q(x) \tag{4.2}$$

BLEU : It analyzes the co-occurrences of n-grams in the ground truth and the proposed responses. It first computes an n-gram precision for the whole dataset.

$$p_n(r,r') = \frac{\sum_k min(h(k,r), h(k,r'_i))}{\sum_k h(k,r_i)} \tag{4.3}$$

where k indexes all possible n-grams of length n and h(k, r) is the number of n-grams k in r. To avoid the drawbacks of using a precision score, namely that it favours shorter (candidate) sentences, the authors introduce a brevity penalty [24].

$$BLEU - N := b(r,r')exp(\sum_{n=1}^{N} \beta_n log P_n(r,r')) \tag{4.4}$$

where β_n is a weighting that is usually uniform, and b is the brevity penalty.

5 RESULTS AND DISCUSSION

The proposed model is trained on aforementioned data-sets and is evaluated using test perplexity and training loss. During training process the learning of model can be validated through performance metrics like loss and perplexity. Performance metrics for proposed work are illustrated in figure , and . The model achieves acceptable test perplexity score as well as BLEU score on test dataset.

Perplexity on test dataset after 140k steps of training is between 90-115 and BLEU score is between 0.7 to 0.9 at same training steps. Though, the perplexity score is little higher than the baseline models for translation proposed model is able to generate acceptable responses at with same checkpoint. This clearly indicates that perplexity and BLEU can be used to evaluate text generation models that solve problem of machine translation same cannot be said for the evaluation of models performing text generation in same language.

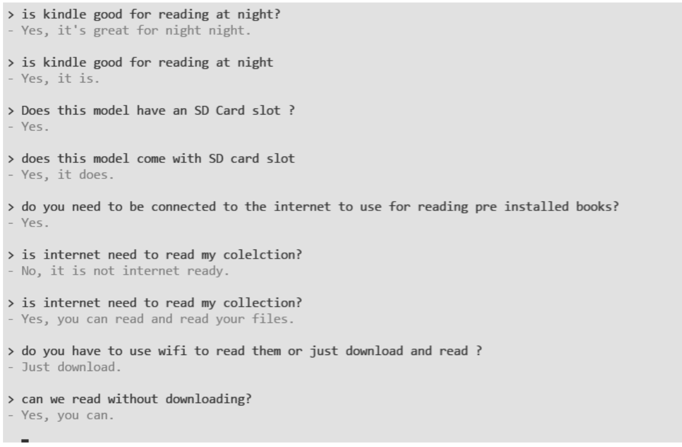


Fig. 3. Sample outputs

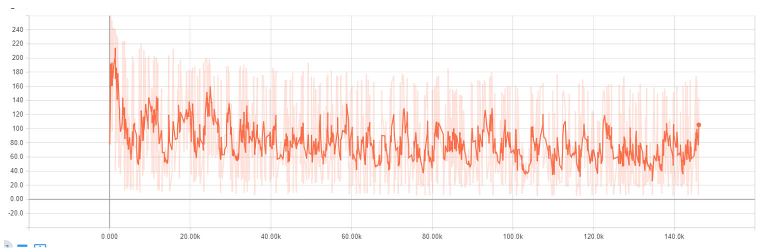


Fig. 4. Training loss

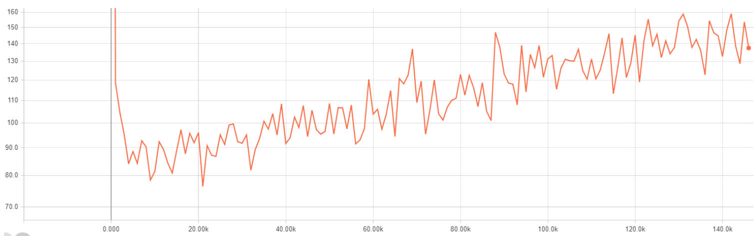


Fig. 5. Test perplexity

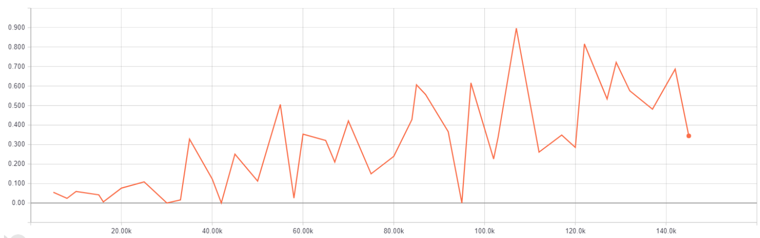


Fig. 6. BLEU score

6 CONCLUSION

Proposed work makes use of the ability of the RNN to map complex structures for the purpose of modeling natural language responses. Results show that the model exhibits the ability to learn relations and dependencies from training data and use the same to generate responses for similar inputs. The model achieves acceptable perplexity and BLEU score on test dataset.

During analysis of results it is observed that over-training results in model getting saturated and generating responses from same set of words regardless of the input. It is also observed that sometimes, model performs better on checkpoints resulting higher perplexity score than those at lower perplexity score.

7 FUTURE ENHANCEMENTS

Proposed approach shows promising results when the inputs given stick to the context, but when intent on conversation changes too rapidly or is out of context which the system can handle the performance degrades. This problem occurs due to absence of flow control mechanism and by providing a action controller kind of mechanism to the system by means of some natural language understanding ability this problem can be better handled.

The performance of system is evaluated on basis of metrics like BLEU and perplexity, but this metrics do not have any context awareness while the system is built for context aware text generation. Since the evaluation metrics do not consider the performance of model based on the words solely rather than the meaning, the numbers indicated by these metrics not necessarily indicate true performance of system. It would be good to evaluate this system with an evaluation parameter which can account for the cohesiveness and meaningfulness of the generated responses with respect to actual input provided.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

References

[1] Sutskever Ilya, Vinyals Oriol, Quoc V Le. Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems. 2014;3104-3112.

[2] Vinyals O, Le Q. A neural conversational model. ArXiv e-prints. Computer Science - Computation and Language. 2015; Available: http://adsabs.harvard.edu/abs/2015 arXiv150605869V (Provided by the SAO/NASA Astrophysics Data System)

[3] Hochreiter Sepp, Schmidhuber Jürgen. Long short-term memory. Neural computation. 1997;9(8):1735-1780.

[4] Cho K, Van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. ArXiv e-prints. Computer Science - Computation and Language, Computer Science - Learning, Computer Science - Neural and Evolutionary Computing, Statistics - Machine Learning. 2014, jun. Available: http://adsabs.harvard.edu/abs/2014ar Xiv1406.1078C, Provided by the SAO/NASA Astrophysics Data System.

[5] Yao K, Zweig G, Peng B. Attention with intention for a neural network conversation model. ArXiv e-prints. Computer Science - Neural and Evolutionary Computing, Computer Science - Artificial Intelligence, Computer Science - Human-Computer Interaction, Computer Science Learning, 2015, oct. Available: http://adsabs.harvard.edu/abs/2015 arXiv151008565Y. Provided by the SAO/NASA Astrophysics Data System.

[6] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Courville, Aaron, Salakhudinov Ruslan, Zemel Rich, Bengio Yoshua. Show, attend and tell: Neural

image caption generation with visual attention. International Conference on Machine Learning. 2015; 2048-2057.

[7] Rush AM, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. ArXiv e-prints; 2015. Available: http://adsabs.harvard.edu/abs/2015 arXiv150900685R

[8] Minh -Thang Luong, Hieu Pham, Christopher D. Manning. Effective approaches to attention-based neural machine translation. CoRR. 2015, Vol abs/1508.04025

[9] Wiseman S, Rush AM. Sequence-to-Sequence Learning as Beam-Search Optimization. ArXiv e-prints, Computer Science - Computation and Language, Computer Science - Learning, Computer Science - Neural and Evolutionary Computing, Statistics - Machine Learning, 2016, jun. Available: http://adsabs.harvard.edu/abs/2016 arXiv160602960W, Provided by the SAO/NASA Astrophysics Data System

[10] Vijayakumar AK, Cogswell M, Selvaraju RR Sun Q, Lee S, Crandall D, Batra D. Diverse beam search: Decoding diverse solutions from neural sequence models. ArXiv e-prints, Computer Science - Artificial Intelligence, Computer Science - Computation and Language, Computer Science - Computer Vision and Pattern Recognition, 2016; oct. Available: http://adsabs.harvard.edu/abs/2016ar Xiv161002424V, Provided by the SAO/NASA Astrophysics Data System.

[11] Sukhbaatar Sainbayar, Weston Jason, Fergus Rob. End-to-end memory networks. Advances in Neural Information Processing Systems. 2015;2440-2448.

[12] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv e-prints, 2014, sep. Available: http://adsabs.harvard.edu/abs/2014 arXiv1409.0473B

[13] Sergey Ioffe, Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning. 2015; 448-456.

[14] Zhang Xiang, Zhao Junbo, LeCun Yann. Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems. 2015;28:649-657. Available: http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf

[15] Chung J, Cho K, BengioY. A character-level decoder without explicit segmentation for neural machine translation. ArXiv e-prints; 2016. Available: http://adsabs.harvard.edu/abs/2016 arXiv160306147C

[16] har-aware-LM, Kim Yoon, Jernite Yacine, Sontag David, Rush Alexander M. Character-aware neural language models. 2016;2741-2749.

[17] Lee J, Cho K, Hofmann T. Fully character-level neural machine translation without explicit segmentation. ArXiv e-prints; 2016. Available: http://adsabs.harvard.edu/abs/2016 arXiv161003017L

[18] Ghosh S, Vinyals O, Strope B, Roy S, Dean T, Heck L. Contextual LSTM (CLSTM) models for large scale NLP tasks. ArXiv e-prints; 2016. Available: http://adsabs.harvard.edu/abs/2016ar Xiv160206291G,

[19] Chung J, Ahn S, Bengio Y. Hierarchical multiscale recurrent neural networks. 2016. Available: http://adsabs.harvard.edu/abs/ 2016arXiv160901704C,

[20] Danescu-Niculescu-Mizil Cristian, Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. ACL 2011.

[21] Julian McAuley. Amazon product data. Available: http://jmcauley.ucsd.edu/data/amazon/, 2017, Online; accessed 30-Nov-2017.

[22] Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. Exploring the limits of language modeling. ArXiv e-prints; 2016. Available: http://adsabs.harvard.edu/abs/2016 arXiv160202410J.

[23] Cross entropy. Available: https://en.m.wikipedia.org/wiki/Cross -entropy, 2017, Online; accessed 30-Nov-2017.

[24] Chia-Wei Liu, Lowe Ryan, V Serban Iulian, Noseworthy Michael, Charlin Laurent, Pineau Joelle. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. ArXiv preprint arXiv:1603.08023, 2016.

Peer-review history:
The peer review history for this paper can be accessed here:
http://prh.sdiarticle3.com/review-history/25635