Generative Deep Neural Networks for Dialogue: A Short Review

Iulian Vlad Serban
Department of Computer Science and Operations Research, University of Montreal

Ryan Lowe
School of Computer Science, McGill University

Laurent Charlin
School of Computer Science, McGill University

Joelle Pineau
School of Computer Science, McGill University

Abstract

Researchers have recently started investigating deep neural networks for dialogue applications. In particular, generative sequence-to-sequence (Seq2Seq) models have shown promising results for unstructured tasks, such as word-level dialogue response generation. The hope is that such models will be able to leverage massive amounts of data to learn meaningful natural language representations and response generation strategies, while requiring a minimum amount of domain knowledge and hand-crafting. An important challenge is to develop models that can effectively incorporate dialogue context and generate meaningful and diverse responses. In support of this goal, we review recently proposed models based on generative encoder-decoder neural network architectures, and show that these models have better ability to incorporate long-term dialogue history, to model uncertainty and ambiguity in dialogue, and to generate responses with high-level compositional structure.

1 Introduction

Researchers have recently started investigating sequence-to-sequence (Seq2Seq) models for dialogue applications. These models typically use neural networks to both represent dialogue histories and to generate or select appropriate responses. Such models are able to leverage large amounts of data in order to learn meaningful natural language representations and generation strategies, while requiring a minimum amount of domain knowledge and hand-crafting. Although the Seq2Seq framework is different from the well-established goal-oriented setting [Gorin et al., 1997, Young, 2000, Singh et al., 2002], these models have already been applied to several real-world applications, with Microsoft's system Xiaoice [Markoff and Mozur, 2015] and Google's Smart Reply system [Kannan et al., 2016] as two prominent examples.

Researchers have mainly explored two types of Seq2Seq models. The first are generative models, which are usually trained with cross-entropy to generate responses word-by-word conditioned on a dialogue context [Ritter et al., 2011, Vinyals and Le, 2015, Sordoni et al., 2015, Shang et al., 2015, Li et al., 2016a, Serban et al., 2016b]. The second are discriminative models, which are trained to select an appropriate response from a set of candidate responses [Lowe et al., 2015, Bordes and Weston, 2016, Inaba and Takahashi, 2016, Yu et al., 2016]. In a related strand of work, researchers have also investigated applying neural networks to the different components of a standard dialogue system, including natural language understanding, natural language generation, dialogue state tracking and

30th Conference on Neural Information Processing Systems (NIPS 2016), Workshop on Learning Methods for Dialogue, Barcelona, Spain.

evaluation [Wen et al., 2016, 2015, Henderson et al., 2013, Mrkšić et al., 2015, Su et al., 2015]. In this paper, we focus on generative models trained with cross-entropy.

One weakness of current generative models is their limited ability to incorporate rich dialogue context and to generate meaningful and diverse responses [Serban et al., 2016b, Li et al., 2016a]. To overcome this challenge, we propose new generative models that are better able to incorporate long-term dialogue history, to model uncertainty and ambiguity in dialogue, and to generate responses with high-level compositional structure. Our experiments demonstrate the importance of the model architecture and the related inductive biases in achieving this improved performance.

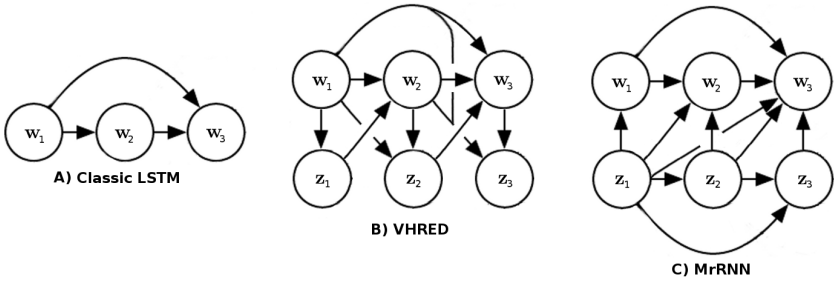

Figure 1: Probabilistic graphical models for dialogue response generation. Variables $\mathbf{w}$ represent natural language utterances. Variables $\mathbf{z}$ represent discrete or continuous stochastic latent variables. (A): Classic LSTM model, which uses a shallow generation process. This is problematic because it has no mechanism for incorporating uncertainty and ambiguity and because it forces the model to generate compositional and long-term structure incrementally on a word-by-word basis. (B): VHRED expands the generation process by adding one latent variable for each utterance, which helps incorporate uncertainty and ambiguity in the representations and generate meaningful, diverse responses. (C): MrRNN expands the generation process by adding a sequence of discrete stochastic variables for each utterance, which helps generate responses with high-level compositional structure.

2 Models

HRED: The Hierarchical Recurrent Encoder-Decoder model (HRED) [Serban et al., 2016b] is a type of Seq2Seq model that decomposes a dialogue into a two-level hierarchy: a sequence of utterances, each of which is a sequence of words. HRED consists of three recurrent neural networks (RNNs): an *encoder* RNN, a *context* RNN and a *decoder* RNN. Each utterance is encoded into a real-valued vector representation by the *encoder* RNN. These utterance representations are given as input to the *context* RNN, which computes a real-valued vector representation summarizing the dialogue at every turn. This summary is given as input to the *decoder* RNN, which generates a response word-by-word. Unlike the RNN encoders in previous Seq2Seq models, the *context* RNN is only updated once every dialogue turn and uses the same parameters for each update. This gives HRED an inductive bias that helps incorporate long-term context and learn invariant representations.

VHRED: The Latent Variable Hierarchical Recurrent Encoder-Decoder model (VHRED) [Serban et al., 2016c] is an HRED model with an additional component: a high-dimensional stochastic latent variable at every dialogue turn. As in HRED, the dialogue context is encoded into a vector representation using *encoder* and *context* RNNs. Conditioned on the summary vector at each dialogue turn, VHRED samples a multivariate Gaussian variable, which is given along with the summary vector as input to the *decoder* RNN. The multivariate Gaussian latent variable allows modelling ambiguity and uncertainty in the dialogue through the latent variable distribution parameters (mean and variance parameters). This provides a useful inductive bias, which helps VHRED encode the dialogue context into a real-valued embedding space even when the dialogue context is ambiguous or uncertain, and it helps VHRED generate more diverse responses.

MrRNN: The Multiresolution RNN (MrRNN) [Serban et al., 2016a] models dialogue as two parallel stochastic sequences: a sequence of high-level coarse tokens (coarse sequences), and a sequence of low-level natural language words (utterances). The coarse sequences follow a latent stochastic process—analogous to hidden Markov models—which conditions the utterances through a hierarchical generation process. The hierarchical generation process first generates the coarse sequence, and conditioned on this generates the natural language utterance. In our experiments, the coarse

Table 1: Model response examples. The arrows indicate change of turn.

Dialogue Context	Model Response
Hello I tried to install ubuntu studio but I get W : Failed to fetch <url > 404 Not Found when apt-get update → the ppa you added is not online	**MrRNN Act. -Ent.:** try a different mirror **MrRNN Noun:** something **VHRED:** but I dont know how to add it **HRED:** how do I do that ?
hi community.. is there any difference between " /.bashrc " " /.bashrc. local " → yes, they're different files :-) the default /.bashrc sources /.bashrc. local → sorry I could undersatnd... If I write on terminal " gedit /.bashrc " it opens a file but how I open " /.bashrc. local " → " gedit /.bashrc. local "... → but it just open blank file nothing inside it	**MrRNN Act.-Ent.:** open up gedit /.bashrc, there is a file called " something " **MrRNN Noun:** the file is empty.. **VHRED:** it's in /.bashrc or /.bashrc **HRED:** what are you trying to do ?
Hey guys what do you in general use for irc something ubuntu xchat or xchat-gnome ? → without -gnome. that is just cut down → you mean drop xchat-gnome and go with xchat ?	**MrRNN Act. -Ent.:** im using xchat right now **MrRNN Noun:** what is xchat-gnome ? **VHRED:** correct **HRED:** yes
when setting up rules with iptables command only writes changes this file " etciptables. rules "? i ask this so i can backup before messing anything → sudo iptables-save something . dat to backup your rules restore with sudo iptables-restore < something . dat	**MrRNN Act. -Ent.:** I don't see any reason why you need iptables to do that **MrRNN Noun:** are you using ubuntu ? **VHRED:** thx **HRED:** thanks

sequences are defined as either noun sequences or activity-entity pairs (predicate-argument pairs) extracted from the natural language utterances. The coarse sequences and utterances are modelled by two separate HRED models. The hierarchical generation provides an important inductive bias, because it helps MrRNN model high-level, compositional structure and generate meaningful and on-topic responses.

3 Experiments

We apply our generative models to dialogue response generation on the Ubuntu Dialogue Corpus [Lowe et al., 2015]. For each example, given a dialogue context, the model must generate an appropriate response. We also present results on Twitter in the Appendix. This task has been studied extensively in the recent literature [Ritter et al., 2011, Sordoni et al., 2015, Li et al., 2016a].

Corpus: The Ubuntu Dialogue Corpus consists of about half a million dialogues extracted from the *#Ubuntu* Internet Relayed Chat (IRC) channel. Users entering this chat channel usually have a specific technical problem. Typically, users first describe their problem, and other users try to help them resolve it. The technical problems range from software-related and hardware-related issues (e.g. installing packages, fixing broken drivers) to informational needs (e.g. finding software).

Evaluation: We carry out an in-lab human study to evaluate the model responses. We recruit 5 human evaluators. We show each evaluator between 30 and 40 dialogue contexts with the ground truth response, and 4 candidate model responses. For each example, we ask the evaluators to compare the candidate responses to the ground truth response and dialogue context, and rate them for fluency and relevancy on a scale 0–4, where 0 means incomprehensible or no relevancy and 4 means flawless English or all relevant. In addition to the human evaluation, we also evaluate dialogue responses w.r.t. the activity-entity metrics proposed by Serban et al. [2016a]. These metrics measure whether the model response contains the same activities (e.g. download, install) and entities (e.g. ubuntu, firefox) as the ground truth responses. Models that generate responses with the same activities and entities as the ground truth responses—including expert responses, which often lead to solving the user's problem—are given higher scores. Sample responses from each model are shown in Table 1.

Table 2: Ubuntu evaluation using F1 metrics w.r.t. activities and entities (mean scores $\pm$ 90% confidence intervals), and human fluency and human relevancy scores given on a scale 0-4 (* indicates scores significantly different from baseline models at 90% confidence)

Model	F1 Activity	F1 Entity	Human Fluency	Human Relevancy
LSTM	1.18 $\pm$0.18	0.87 $\pm$0.15	-	-
HRED	4.34 $\pm$0.34	2.22 $\pm$0.25	2.98	1.01
VHRED	4.63 $\pm$0.34	2.53 $\pm$0.26	-	-
MrRNN Noun	4.04 $\pm$0.33	**6.31 $\pm$0.42**	**3.48***	**1.32***
MrRNN Act.-Ent.	**11.43 $\pm$0.54**	3.72 $\pm$0.33	**3.42***	1.04

Results: The results are given in Table 2. The MrRNNs perform substantially better than the other models w.r.t. both the human evaluation study and the evaluation metrics based on activities and

entities. MrRNN with noun representations obtains an F1 entity score at 6.31, while all other models obtain less than half F1 scores between $0.87 - 2.53$, and human evaluators consistently rate its fluency and relevancy significantly higher than all the baseline models. MrRNN with activity representations obtains an F1 activity score at 11.43, while all other models obtain less than half F1 activity scores between $1.18 - 4.63$, and performs substantially better than the baseline models w.r.t. the F1 entity score. This indicates that the MrRNNs have learned to model high-level, goal-oriented sequential structure in the Ubuntu domain. Followed by these, VHRED performs better than the HRED and LSTM models w.r.t. both activities and entities. This shows that VHRED generates more appropriate responses, which suggests that the latent variables are useful for modeling uncertainty and ambiguity. Finally, HRED performs better than the LSTM baseline w.r.t. both activities and entities, which underlines the importance of representing longer-term context. These conclusions are confirmed by additional experiments on response generation for the Twitter domain (see Appendix).

4 Discussion

We have presented generative models for dialogue response generation. We have proposed architectural modifications with inductive biases towards 1) incorporating longer-term context, 2) handling uncertainty and ambiguity, and 3) generating diverse and on-topic responses with high-level compositional structure. Our experiments show the advantage of the architectural modifications quantitatively through human experiments and qualitatively through manual inspections. These experiments demonstrate the need for further research into generative model architectures. Although we have focused on three generative models, other model architectures such as memory-based models [Bordes and Weston, 2016, Weston et al., 2015] and attention-based models [Shang et al., 2015] have also demonstrated promising results and therefore deserve the attention of future research.

In another line of work, researchers have started proposing alternative training and response selection criteria [Weston, 2016]. Li et al. [2016a] propose ranking candidate responses according to a mutual information criterion, in order to incorporate dialogue context efficiently and retrieve on-topic responses. Li et al. [2016b] further propose a model trained using reinforcement learning to optimize a hand-crafted reward function. Both these models are motivated by the lack of *diversity* observed in the generative model responses. Similarly, Yu et al. [2016] propose a hybrid model—combining retrieval models, neural networks and hand-crafted rules—trained using reinforcement learning to optimize a hand-crafted reward function. In contrast to these approaches, without combining several models or having to modify the training or response selection criterion, VHRED generates more diverse responses than previous models. Similarly, by optimizing the joint log-likelihood over sequences, MrRNNs generate more appropriate and on-topic responses with compositional structure. Thus, improving generative model architectures has the potential to compensate — or even remove the need — for hand-crafted reward functions.

At the same time, the models we propose are not necessarily better language models, which are more efficient at compressing dialogue data as measured by word perplexity. Although these models produce responses that are preferred by humans, they often result in higher test set perplexity than traditional LSTM language models. This suggests maximizing log-likelihood (i.e. minimizing perplexity) is not a sufficient training objective for these models. An important line of future work therefore lies in improving the objective functions for training and response selection, as well as learning directly from interactions with real users.

References

A. Bordes and J. Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.

A. L. Gorin, G. Riccardi, and J. H. Wright. How may i help you? *Speech communication*, 23(1):113–127, 1997.

M. Henderson, B. Thomson, and S. Young. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 467–471, 2013.

M. Inaba and K. Takahashi. Neural utterance ranking model for conversational dialogue systems. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 393, 2016.

A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukács, M. Ganea, P. Young, et al. Smart reply: Automated response suggestion for email. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, volume 36, pages 495–503, 2016.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL*, 2016a.

J. Li, W. Monroe, A. Ritter, and D. Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016b.

R. Lowe, N. Pow, I. Serban, and J. Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proc. of SIGDIAL-2015*, 2015.

J. Markoff and P. Mozur. For sympathetic ear, more chinese turn to smartphone program. *NY Times*, 2015.

N. Mrkšić, D. O. Séaghdha, B. Thomson, M. Gašić, P.-H. Su, D. Vandyke, T.-H. Wen, and S. Young. Multi-domain dialog state tracking using recurrent neural networks. In *HLT-NAACL*, pages 120–129, 2015.

A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *EMNLP*, 2011.

I. V. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. *arXiv preprint arXiv:1606.00776*, 2016a.

I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784, 2016b.

I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*, 2016c.

L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. In *ACL-IJCNLP*, pages 1577–1586, 2015.

S. Singh, D. Litman, M. Kearns, and M. Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *JAIR*, 16:105–133, 2002.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2015)*, 2015.

P.-H. Su, D. Vandyke, M. Gasic, D. Kim, N. Mrksic, T.-H. Wen, and S. Young. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. In *SIGDIAL*, 2015.

O. Vinyals and Q. Le. A neural conversational model. *ICML, Workshop*, 2015.

T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1199`.

T.-H. Wen, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, D. Vandyke, and S. Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv:1604.04562*, 2016.

J. Weston. Dialog-based language learning. *arXiv preprint arXiv:1604.06045*, 2016.

J. Weston, S. Chopra, and A. Bordes. Memory networks. *ICLR*, 2015.

S. Young. Probabilistic methods in spoken–dialogue systems. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769), 2000.

Z. Yu, Z. Xu, A. W. Black, and A. I. Rudnicky. Strategy and policy learning for non-task-oriented conversational systems. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 404, 2016.

Appendix

Twitter Results

Corpus: We experiment on a Twitter Dialogue Corpus [Ritter et al., 2011] containing about one million dialogues. The task is to generate utterances to append to existing Twitter conversations. This task is typically categorized as a non-goal-driven task, because any fluent and on-topic response may be adequate.

Evaluation: We carry out a human study on Amazon Mechanical Turk (AMT). We show human evaluators a dialogue context along with two potential responses: one response generated from each model conditioned on the dialogue context. We ask evaluators to choose the response most appropriate to the dialogue context. If the evaluators are indifferent, they can choose neither response. For each pair of models we conduct two experiments: one where the example contexts contain at least 80 unique tokens (*long context*), and one where they contain at least 20 (not necessarily unique) tokens (*short context*). We experiment with the LSTM, HRED and VHRED models, as well as a TF-IDF retrieval-based baseline model. We do not experiment with the MrRNN models, because we do not have appropriate coarse representations for this domain.

Results: The results given in Table 3 show that VHRED is strongly preferred in the majority of the experiments. In particular, VHRED is strongly preferred over the HRED and TF-IDF baseline models for both short and long context settings. VHRED is also strongly preferred over the LSTM baseline model for long contexts, although the LSTM model is preferred over VHRED for short contexts.For short contexts, the LSTM model is often preferred over VHRED because the LSTM model tends to generate very *generic* responses. Such *generic* or *safe* responses are reasonable for a wide range of contexts, but are not useful when applied through-out a dialogue, because the user would loose interest in the conversation.

In conclusion, VHRED performs substantially better overall than competing models, which suggests that the high-dimensional latent variables help model uncertainty and ambiguity in the dialogue context and help generate meaningful responses.

Table 3: Wins, losses and ties (in %) of VHRED against baselines based on the human study (mean preferences $\pm$ 90% confidence intervals, where $*$ indicates significant differences at 90% confidence)

Opponent	Wins	Losses	Ties
Short Contexts			
VHRED vs LSTM	32.3 ± 2.4	$\mathbf{42.5 \pm 2.6}^*$	25.2 ± 2.3
VHRED vs HRED	$\mathbf{42.0 \pm 2.8}^*$	31.9 ± 2.6	26.2 ± 2.5
VHRED vs TF-IDF	$\mathbf{51.6 \pm 3.3}^*$	17.9 ± 2.5	30.4 ± 3.0
Long Contexts			
VHRED vs LSTM	$\mathbf{41.9 \pm 2.2}^*$	36.8 ± 2.2	21.3 ± 1.9
VHRED vs HRED	$\mathbf{41.5 \pm 2.8}^*$	29.4 ± 2.6	29.1 ± 2.6
VHRED vs TF-IDF	$\mathbf{47.9 \pm 3.4}^*$	11.7 ± 2.2	40.3 ± 3.4