

Coherent Dialogue with Attention-based Language Models

Hongyuan Mei
Johns Hopkins University
hmei@cs.jhu.edu

Mohit Bansal
UNC Chapel Hill
mbansal@cs.unc.edu

Matthew R. Walter
TTI-Chicago
mwalter@ttic.edu

Abstract

We model coherent conversation continuation via RNN-based dialogue models equipped with a dynamic attention mechanism. Our attention-RNN language model dynamically increases the scope of attention on the history as the conversation continues, as opposed to standard attention (or alignment) models with a fixed input scope in a sequence-to-sequence model. This allows each generated word to be associated with the most relevant words in its corresponding conversation history. We evaluate the model on two popular dialogue datasets, the open-domain *MovieTriples* dataset and the closed-domain *Ubuntu Troubleshoot* dataset, and achieve significant improvements over the state-of-the-art and baselines on several metrics, including complementary diversity-based metrics, human evaluation, and qualitative visualizations. We also show that a vanilla RNN with dynamic attention outperforms more complex memory models (e.g., LSTM and GRU) by allowing for flexible, long-distance memory. We promote further coherence via topic modeling-based reranking.

Introduction

Automatic conversational models (Winograd 1971), also known as dialogue systems, are of great importance to a large variety of applications, ranging from open-domain entertaining chatbots to goal-oriented technical support agents. An increasing amount of research has recently been done to build purely data-driven dialogue systems that learn from large corpora of human-to-human conversations, without using hand-crafted rules or templates. While most work in this area formulates dialogue modeling in a sequence-to-sequence framework (similar to machine translation) (Ritter, Cherry, and Dolan 2011; Shang, Lu, and Li 2015; Vinyals and Le 2015; Sordani et al. 2015; Li et al. 2016a; Dušek and Jurčiček 2016), some more recent work (Serban et al. 2016; Luan, Ji, and Ostendorf 2016) instead trains a language model over the entire dialogue as one single sequence. In our work, we empirically demonstrate that a language model is better suited to dialogue modeling, as it learns how the conversation evolves as information progresses. Sequence-to-sequence models, on the other hand, learn only how the most recent dialogue response is generated. Such models are better suited to converting the same

information from one modality to another, e.g., in machine translation and image captioning.

We improve the coherence of such neural dialogue language models by developing a generative dynamic attention mechanism that allows each generated word to choose which related words it wants to align to in the increasing conversation history (including the previous words in the response being generated). Neural attention (or alignment) has proven very successful for various sequence-to-sequence tasks by associating salient items in the source sequence with the generated item in the target sequence (Mnih et al. 2014; Bahdanau, Cho, and Bengio 2015; Xu et al. 2015; Mei, Bansal, and Walter 2016a; Parikh et al. 2016). However, such attention models are limited to a fixed scope of history, corresponding to the input source sequence. In contrast, we introduce a dynamic attention mechanism to a recurrent neural network (RNN) language model in which the scope of attention increases as the recurrence operation progresses from the start through the end of the conversation.

The dynamic attention model promotes coherence of the generated dialogue responses (continuations) by favoring the generation of words that have syntactic or semantic associations with salient words in the conversation history. Our simple model shows significant improvements over state-of-the-art models and baselines on several metrics (including complementary diversity-based metrics, human evaluation, and qualitative visualizations) for two benchmark datasets, the open-domain *MovieTriples* and closed-domain *Ubuntu Troubleshoot* datasets. Our vanilla RNN model with dynamic attention outperforms more complex memory models (e.g., LSTM and GRU) by allowing for long-distance and flexible memory. We also present several visualizations to intuitively understand what the attention model is learning. Finally, we also explore a complementary LDA-based method to re-rank the outputs of the soft alignment-based coherence method, further improving performance on the evaluation benchmarks.

Related Work

A great deal of attention has been paid to developing data-driven methods for natural language dialogue generation. Conventional statistical approaches tend to rely extensively on hand-crafted rules and templates, require interaction with

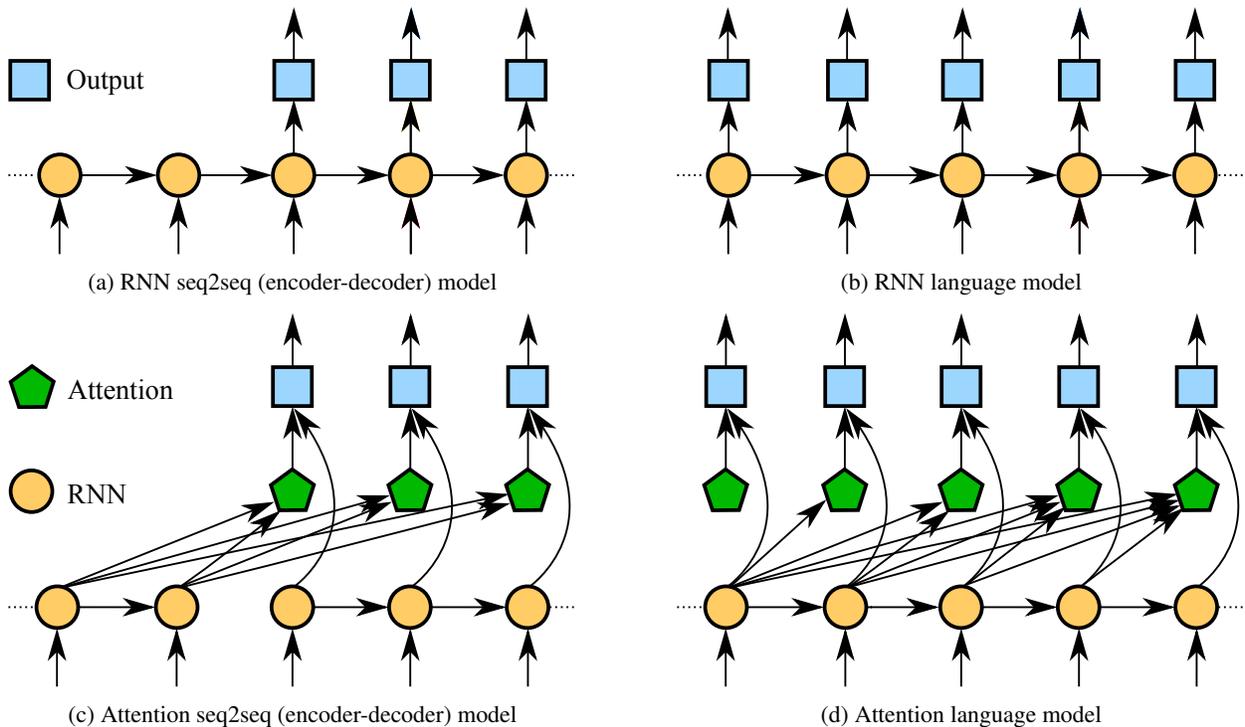


Figure 1: Comparing RNN language models to RNN sequence-to-sequence model, with and without attention.

humans or simulated users to optimize parameters, or produce conversation responses in an information retrieval fashion. Such properties prevent training on the large human conversational corpora that are becoming increasingly available, or fail to produce novel natural language responses.

Ritter, Cherry, and Dolan (2011) formulate dialogue response generation as a statistical phrase-based machine translation problem, which requires no explicit hand-crafted rules. The recent success of RNNs in statistical machine translation (Sutskever, Vinyals, and Lee 2014; Bahdanau, Cho, and Bengio 2015) has inspired the application of such models to the field of dialogue modeling. Vinyals and Le (2015) and Shang, Lu, and Li (2015) employ an RNN to generate responses in human-to-human conversations by treating the conversation history as one single temporally ordered sequence. In such models, the distant relevant context in the history is difficult to recall. Some efforts have been made to overcome this limitation. Sordani et al. (2015) separately encode the most recent message and all the previous context using a bag-of-words representation, which is decoded using an RNN. This approach equates the distance of each word in the generated output to all the words in the conversation history, but loses the temporal information of the history. Serban et al. (2016) design a hierarchical model that stacks an utterance-level RNN on a token-level RNN, where the utterance-level RNN reduces the number of computational steps between utterances. Wen et al. (2015) and Wen et al. (2016) improve spoken dialog systems via multi-domain and semantically conditioned neural networks on dialog act representations and explicit slot-value formulations.

Our work explores the ability of recurrent neural network language models (Bengio et al. 2003; Mikolov 2010) to interpret and generate natural language conversations while still maintaining a relatively simple architecture. We show that a language model approach outperforms the sequence-to-sequence model at dialogue modeling. Recently, Tran, Bisazza, and Monz (2016) demonstrated that the neural attention mechanism can improve the effectiveness of a neural language model. We propose an attention-based neural language model for dialogue modeling that learns how a conversation evolves as a whole, rather than only how the most recent response is generated, and that also reduces the number of computations between the current recurrence step and the distant relevant context in the conversation history.

The attention mechanism in our model has the additional benefit of favoring words that have semantic association with salient words in the conversation history, which promotes the *coherence* of the topics in the continued dialogue. This is important when conversation participants inherently want to maintain the topic of the discussion. Some past studies have equated coherence with propositional consistency (Goldberg 1983), while others see it as a summary impression (Sanders 1983). Our work falls in the category of viewing coherence as topic continuity (Crow 1983; Sigman 1983). Similar objectives, i.e., generating dialogue responses with certain properties, have been addressed recently, such as promoting response diversity (Li et al. 2016a), enhancing personal consistency (Li et al. 2016b), and improving specificity (Yao et al. 2016). Concurrent with this work, Luan, Ji, and Ostendorf (2016) improve topic

consistency by feeding into the model the learned LDA-based topic representations. We show that the simple attention neural language model significantly outperforms such a design. Furthermore, we suggest an LDA-based re-ranker complementary to soft neural attention that further promotes topic coherence.

The Model

In this section, we introduce an attention-RNN dialogue model and compare it to the architectures of basic RNN and sequence-to-sequence models.

RNN Seq2Seq and Language Models

Recurrent neural networks have been successfully used both in sequence-to-sequence models (RNN-Seq2Seq, Fig. 1a) (Sutskever, Vinyals, and Lee 2014) and in language models (RNN-LM, Fig. 1b) (Bengio et al. 2003; Mikolov 2010). We first discuss language models for dialogue, which is the primary focus of our work, then briefly introduce the sequence-to-sequence model, and lastly discuss the use of attention methods in both models.

The RNN-LM models a sentence as a sequence of tokens $\{w_0, w_1, \dots, w_T\}$ with a recurrence function

$$h_t = f(h_{t-1}, w_{t-1}) \quad (1)$$

and an output (softmax) function

$$P(w_t = v_j | w_{0:t-1}) = \frac{\exp g(h_t, v_j)}{\sum_i \exp g(h_t, v_i)}, \quad (2)$$

where the recurrent hidden state $h_t \in \mathbb{R}^d$ encodes all the tokens up to $t-1$ and is used to compute the probability of generating $v_j \in V$ as the next token from the vocabulary V .

The functions f and g are typically defined as

$$f(h_{t-1}, w_{t-1}) = \tanh(Hh_{t-1} + PE_{w_{t-1}}) \quad (3a)$$

$$g(h_t, v_j) = O_{v_j}^\top h_t, \quad (3b)$$

where $H \in \mathbb{R}^{d \times d}$ is the recurrence matrix, $E_{w_{t-1}}$ is a column of word embedding matrix $E \in \mathbb{R}^{d_e \times V}$ that corresponds to w_{t-1} , $P \in \mathbb{R}^{d \times d_e}$ projects word embedding into the space of the same dimension d as the hidden units, and $O \in \mathbb{R}^{d \times V}$ is the output word embedding matrix with column vector O_{v_j} corresponding to v_j .

We train the RNN-LM, i.e. estimate the parameters H , P , E and O , by maximizing the log-likelihood on a set of natural language training sentences of size N

$$\ell = \frac{1}{N} \sum_{n=1}^N \sum_{t=0}^{T_n} \log P(w_t | w_{0:t-1}) \quad (4)$$

Since the entire architecture is differentiable, the objective can be optimized by back-propagation.

When dialogue is formulated as a sequence-to-sequence task, the RNN-Seq2Seq model can be used in order to predict a target sequence $w_{0:L}^T = \{w_0^T, w_1^T, \dots, w_L^T\}$ given an input source sequence $w_{0:M}^S = \{w_0^S, w_1^S, \dots, w_M^S\}$. In such settings, an encoder RNN represents the input as a sequence of hidden states $h_{0:M}^S = \{h_0^S, h_1^S, \dots, h_M^S\}$, and a separate decoder RNN then predicts the target sequence token-by-token given the encoder hidden states $h_{0:M}^S$.

Attention in RNN-Seq2Seq Models

There are several ways by which to integrate the sequence of hidden states $h_{0:M}^S$ in the decoder RNN. An attention mechanism (Fig. 1c) has proven to be particularly effective for various related tasks in machine translation, image caption synthesis, and language understanding (Mnih et al. 2014; Bahdanau, Cho, and Bengio 2015; Xu et al. 2015; Mei, Bansal, and Walter 2016a).

The attention module takes as input the encoder hidden state sequence $h_{0:M}^S$ and the decoder hidden state h_{l-1}^T at each step $l-1$, and returns a context vector z_l computed as a weighted average of encoder hidden states $h_{0:M}^S$

$$\beta_{lm} = b^\top \tanh(Wh_{l-1}^T + Uh_m^S) \quad (5a)$$

$$\alpha_{lm} = \exp(\beta_{lm}) / \sum_{m=0}^M \exp(\beta_{lm}) \quad (5b)$$

$$z_l = \sum_{m=0}^M \alpha_{lm} h_m^S, \quad (5c)$$

where parameters $W \in \mathbb{R}^{d \times d}$, $U \in \mathbb{R}^{d \times d}$, and $b \in \mathbb{R}^d$ are jointly learned with the other model parameters. The context vector z_l is then used as an extra input to the decoder RNN at step l together with $w_{0:l-1}^T$ to predict the next token w_l^T .

Attention in RNN-LM

We develop an attention-RNN language model (A-RNN-LM) as illustrated in Figure 1d, and describe how it can be used in the context of dialogue modeling. We then describe its advantages compared to the use of attention in sequence-to-sequence models.

As with the RNN-LM, the model first encodes the input into a sequence of hidden states up to word $t-1$ (Eqn. 1). Given a representation of tokens up to $t-1$ $\{r_0, r_1, \dots, r_{t-1}\}$ (which we define shortly), the attention module computes the context vector z_t at step t as a weighted average of $r_{0:t-1}$

$$\beta_{ti} = b^\top \tanh(Wh_{t-1} + Ur_i) \quad (6a)$$

$$\alpha_{ti} = \exp(\beta_{ti}) / \sum_{i=0}^{t-1} \exp(\beta_{ti}) \quad (6b)$$

$$z_t = \sum_{i=0}^{t-1} \alpha_{ti} r_i \quad (6c)$$

We then use the context vector z_t together with the hidden state h_t to predict the output at time t

$$g(h_t, z_t, v_j) = O_{v_j}^\top (O_h h_t + O_z z_t) \quad (7a)$$

$$P(w_t = v_j | w_{0:t-1}) = \frac{\exp g(h_t, z_t, v_j)}{\sum_i \exp g(h_t, z_t, v_i)}, \quad (7b)$$

where $O_h \in \mathbb{R}^{d \times d}$ and $O_z \in \mathbb{R}^{d \times d_z}$ project h_t and z_t , respectively, into the same space of dimension d .

There are multiple benefits of using an attention-RNN language model for dialogue, which are empirically supported by our experimental results. First, a complete dialogue is usually composed of multiple turns. A language

model over the entire dialogue is expected to better learn how a conversation evolves as a whole, unlike a sequence-to-sequence model, which only learns how the most recent response is generated and is better suited to translation-style tasks that transform the same information from one modality to another. Second, compared to LSTM models, an attention-based RNN-LM also allows for gapped context and a flexible combination of conversation history for every individual generated token, while maintaining low model complexity. Third, attention models yield interpretable results—we visualize the learned attention weights, showing how attention chooses the salient words from the dialogue history that are important for generating each new word. Such a visualization is typically harder for the hidden states and gates of conventional LSTM and RNN language models.

With an attention mechanism, there are multiple options for defining the token representation $r_{0:t-1}$. The original attention model introduced by Bahdanau, Cho, and Bengio (2015) uses the hidden units $h_{0:t-1}$ as the token representations $r_{0:t-1}$. Recent work (Mei, Bansal, and Walter 2016a; 2016b) has demonstrated that performance can be improved by using multiple abstractions of the input, e.g., $r_i = (E_{w_i}^\top, h_i^\top)^\top$, which is what we use in this work.

LDA-based Re-Ranking

While the trained attention-RNN dialogue model generates a natural language continuation of a conversation while maintaining topic concentration by token association, some dialogue-level topic-supervision can help to encourage generations that are more topic-aware. Such supervision is not commonly available, and we use unsupervised methods to learn document-level latent topics. We employ the learned topic model to select the best continuation based on document-level topic-matching.

We choose Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003; Blei and Lafferty 2009) due to its demonstrated ability to learn a distribution over latent topics given a collection of documents. This generative model assumes documents $\{w_{0:T_n}\}_{n=1}^N$ arise from K topics, each of which is defined as a distribution over a fixed vocabulary of terms, which forms a graphical structure \mathcal{L} that can be learned from the training data. The topic representation $\hat{\theta}$ of a (possibly unseen) dialogue $w_{0:T}$ can then be estimated with the learned topic structure \mathcal{L} as $\hat{\theta}(w_{0:T}) = \mathcal{L}(w_{0:T})$.

Given a set of generated continuations $\{c_m\}_{m=1}^M$ for each unseen dialogue $w_{0:T}$, the topic representations of the dialogue and its continuations are $\hat{\theta}(w_{0:T}) = \mathcal{L}(w_{0:T})$ and $\hat{\theta}(c_m) = \mathcal{L}(c_m)$, respectively. We employ a matching score $\mathcal{S}_m = \mathcal{S}(\hat{\theta}(w_{0:T}), \hat{\theta}(c_m))$ to compute the similarity between $\hat{\theta}(w_{0:T})$ and each $\hat{\theta}(c_m)$. In the end, a weighted score is computed as $\tilde{\mathcal{S}}_m = \lambda \mathcal{S}_m + (1 - \lambda) \ell(c_m | w_{0:T})$, where $\lambda \in [0, 1]$ and $\ell(c_m | w_{0:T})$ is the conditional log-likelihood of the continuation c_m . The hyper-parameters K and λ are tuned on a development set.

Concurrent with our work, Luan, Ji, and Ostendorf (2016) use learned topic representations $\hat{\theta}$ of the given conversation

as an extra feature in a language model to enhance the topic coherence of the generation. As we show in the Results and Analysis section, our model significantly outperforms this approach.

Experimental Setup

Dataset

We train and evaluate the models on two large natural language dialogue datasets, *MovieTriples* (pre-processed by Serban et al. (2016)) and *Ubuntu Troubleshoot* (pre-processed by Luan, Ji, and Ostendorf (2016)). The dialogue within each of these datasets consists of a sequence of utterances (turns), each of which is a sequence of tokens (words).¹ The accompanying supplementary material provides the statistics for these two datasets.

Evaluation Metrics

For the sake of comparison, we closely follow previous work and adopt several standard (and complementary) evaluation metrics: perplexity (PPL), word error rate (WER), recall@N, BLEU, and diversity-based Distinct-1. We provide further discussion of the various metrics and their advantages in the supplementary material. On the *MovieTriples* dataset, we use PPL and WER, as is done in previous work. Following Serban et al. (2016), we adopt two versions for each metric: i) PPL as the word-level perplexity over the entire dialogue conversation; ii) PPL@L as the word-level perplexity over the last utterance of the conversation; iii) WER; and iv) WER@L (defined similarly).

On the *Ubuntu* dataset, we follow previous work and use PPL and recall@N. Recall@N (Manning et al. 2008) evaluates a model by measuring how often the model ranks the correct dialogue continuation within top-N given 10 candidates. Additionally, we also employ the BLEU score (Papineni et al. 2001) to evaluate the quality of the generations produced by the models. Following Luan, Ji, and Ostendorf (2016), we perform model selection using PPL on the development set, and perform the evaluation on the test set using the other metrics. We also present evaluation using the Distinct-1 metric (proposed by Li et al. (2016a)) to measure the ability of the A-RNN to promote diversity in the generations, because typical neural dialogue models generate generic, safe responses (technically appropriate but not informative, e.g., “I dont know”). Finally, we also present a preliminary human evaluation.

Training Details

For the *MovieTriples* dataset, we follow the same procedure as Serban et al. (2016) and first pretrain on the large *Q-A SubTitle* dataset (Ameixa et al. 2014), which contains 5.5M question-answer pairs from which we randomly sample 20000 pairs as the held-out set, and then fine-tune on the target *MovieTriples* dataset. We perform early-stopping according to the PPL score on the held-out set. We train the

¹Following Luan, Ji, and Ostendorf (2016), we randomly sample nine utterances as negative examples of the last utterance for each conversation in *Ubuntu Troubleshoot* for the development set.

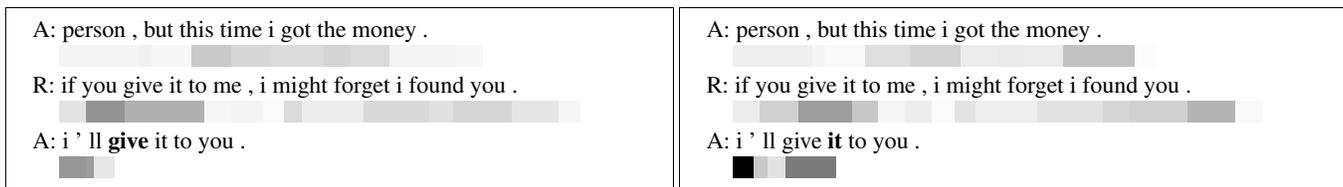
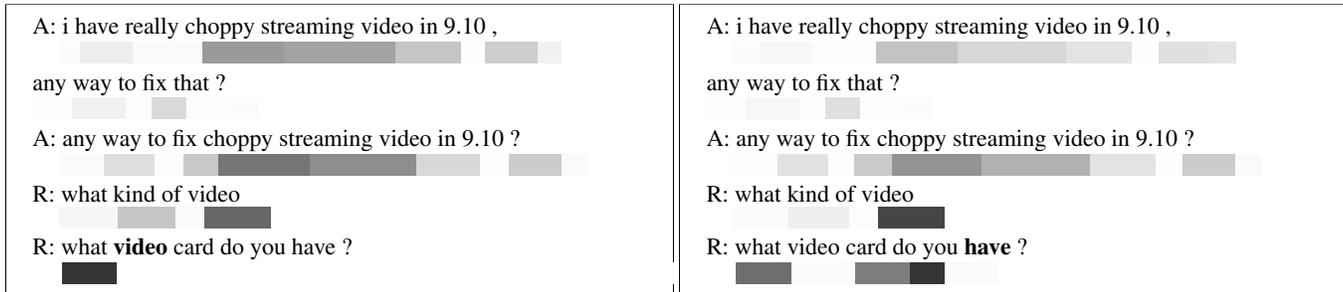
(a) *MovieTriples*(b) *Ubuntu Troubleshoot*

Figure 2: A visualization of attention on the (a) *MovieTriples* and (b) *Ubuntu Troubleshooting* datasets, showing which words in the conversation history are being aligned to, for each generated response word. Shaded intervals indicate the strength with which the corresponding words in the conversation history and response are attend to when generating the bolded word in the response. We show this for two generated words in the same response (left and right column).

models for both the *MovieTriples* and *Ubuntu Troubleshoot* datasets using Adam (Kingma and Ba 2015) for optimization in RNN back-propagation. The accompanying supplementary material provides additional training details, including the hyper-parameter settings.

Results and Analysis

Primary Dialogue Modeling Results

In this section, we compare the performance on several metrics of our attention-based RNN-LM with RNN baselines and state-of-the-art models on the two benchmark datasets. Table 1 reports PPL and WER results on the *MovieTriples* test set, while Table 2 compares different models on *Ubuntu Troubleshoot* in terms of PPL on the development set and recall@N ($N = 1$ and 2) on the test set (following what previous work reports). In the tables, RNN is the plain vanilla RNN language model (RNN-LM), as defined in The Model section, and LSTM is an LSTM-RNN language model, i.e., an RNN-LM with LSTM memory units. A-RNN refers to our main model as defined in the Attention in RNN-LM section. HRED in Table 1 is the hierarchical neural dialogue model proposed by Serban et al. (2016).² LDA-CONV in Table 2 is proposed by Luan, Ji, and Ostendorf (2016), which integrates learned LDA-topic-proportions into an LSTM language model in order to promote topic-concentration in the generations. Both tables demonstrate that the attention-RNN-LM (A-RNN) model achieves the best results reported to-date on these datasets in terms all evaluation metrics. It improves the ability of

²We compare to their best-performing model version, that adopts bidirectional gated-unit RNN (GRU).

Table 1: Results on the *MovieTriples* test set. The HRED results are from Serban et al. (2016).

Model	PPL	PPL@L	WER	WER@L
RNN	27.09	26.67	64.10%	64.07%
HRED	26.81	26.31	63.93%	63.91%
A-RNN	25.52	23.46	61.58%	60.15%

Table 2: *Ubuntu Troubleshoot* PPL and recall@N, with LSTM and LDA-CONV results from Luan et al. (2016).

Model	PPL	recall@1	recall@2
RNN	56.16	11%	22%
LSTM	54.93	12%	22%
LDA-CONV	51.13	13%	24%
A-RNN	45.38	17%	30%

an RNN-LM to model continuous dialogue conversations, while keeping the model architecture simple.

Table 3: RNN-LM vs. RNN-Seq2Seq

PPL	<i>MovieTriples</i>	<i>Ubuntu Troubleshoot</i>
RNN-Seq2Seq	35.10	104.61
RNN-LM	27.81	56.16

We also evaluate the effectiveness of the RNN-LM and RNN-Seq2Seq models on both the *MovieTriples* and *Ubuntu Troubleshoot* development sets. As shown in Table 3, the RNN language model yields lower perplexity than the RNN

Table 4: BLEU score on *Ubuntu Troubleshoot*

Model	dev BLEU	test BLEU
RNN	0.1846	0.1692
A-RNN	0.2702	0.3713
T-A-RNN	0.2908	0.3128
A-RNN-RR	0.4696	0.4279
T-A-RNN-RR	0.4895	0.3971

History: A: person , but this time i got the money .
R: if you give it to me , i might forget i found you .

A-RNN: A: i ' ll give it to you .
RNN: A: i don ' t know .
Reference: A: i don ' t have it with me . tell person .

(a) *MovieTriples*

History: A: i have really choppy streaming video in 9.10, any way to fix that?
A: any way to fix choppy streaming video in 9.10 ?
R: what kind of video

A-RNN: A: what video card do you have ?
RNN: A: what are you trying to do ?
Reference: A: what cpu do you have

(b) *Ubuntu Troubleshoot*

Figure 3: Examples of responses generated by the A-RNN.

sequence-to-sequence model on both datasets. Hence, we present all primary results on our primary A-RNN attention-based RNN language model.³

Generation Diversity Results

Next, we investigate the ability of the A-RNN to promote diversity in the generations, compared to that of the vanilla RNN using the Distinct-1 metric proposed by Li et al. (2016a). Distinct-1 is computed as the number of distinct unigrams in the generation scaled by the total number of generated tokens. Table 5 shows that our attention-based RNN language model (A-RNN) yields much more diversity in its generations as compared to the vanilla RNN baseline.

Topic Coherence Results

Next, we investigate the ability of the different models to promote topic coherence in the generations in terms of BLEU score. In addition to the RNN and A-RNN models, we consider T-A-RNN, a method that incorporates LDA-based topic information into an A-RNN model, following the approach of Luan, Ji, and Ostendorf (2016). We also evaluate our LDA-based re-ranker, A-RNN-RR, which re-ranks according to the score $\tilde{S}_m = \lambda S_m + (1 - \lambda)\ell(c_m|w_{0:T})$, where we compute the log-likelihood $\ell(c_m|w_{0:T})$ based upon a trained A-RNN-M model and validate the weight λ on the development set. We also consider a method that combines the T-A-RNN model

³Experiments also demonstrate significant improvements for the Attention-RNN-LM over the Attention-RNN-Seq2Seq.

Table 5: Generation Diversity Results: A-RNN vs. RNN

Distinct-1	<i>MovieTriples</i>	<i>Ubuntu Troubleshoot</i>
RNN	0.0004	0.0007
A-RNN	0.0028	0.0104

with an LDA-based re-ranker (T-A-RNN-RR).⁴ Table 4 reports the resulting BLEU scores for each of these methods on the development and test sets from the *Ubuntu Troubleshoot* dataset. We make the following observations based upon these results: (1) The A-RNN performs substantially better than the RNN with regards to BLEU; (2) using our LDA-based re-ranker further improves the performance by a significant amount (A-RNN v.s. A-RNN-RR); (3) as opposed to our LDA-based re-ranker, adopting the LDA design of Luan, Ji, and Ostendorf (2016) only yields marginal improvements on the development set, but does not generalize well to the test set (A-RNN v.s. T-A-RNN and A-RNN-RR v.s. T-A-RNN-RR). Also, our LDA re-ranker results in substantial improvements even on top of their topic-based model (T-A-RNN v.s. T-A-RNN-RR).

Preliminary Human Evaluation

In addition to multiple automatic metrics, we also report a preliminary human evaluation. On each dataset, we manually evaluate the generations of both the A-RNN and RNN models on 100 examples randomly sampled from the test set. For each example, we randomly shuffle the two response generations, anonymize the model identity, and ask a human annotator to choose which response generation is more topically coherent based on the conversation history. As Table 6 shows, the A-RNN model wins substantially more often than the RNN model.

Table 6: Human Evaluation: A-RNN vs. RNN

	<i>MovieTriples</i>	<i>Ubuntu Troubleshoot</i>
Not distinguishable	48%	74%
RNN wins	6%	5%
A-RNN wins	46%	21%

Qualitative Analysis

Next, we qualitatively evaluate the effectiveness of our A-RNN model through visualizations of the attention and outputs on both datasets. Figure 2 provides a visualization of the attention for a subset of the words in the generation for the two datasets. The last line in both Fig. 2a and Fig. 2b presents the generated response and we highlight in bold two output words (one on the left and one on the right) for two time steps. For each highlighted generated word, we visualize the attention weights for words in the conversation history (i.e., words in the preceding turns and those previously

⁴Since Luan, Ji, and Ostendorf (2016) do not publish BLEU scores or implementations of their models, we can not compare with LDA-CONV on BLEU. Instead, we demonstrate the effect of adding the key component of LDA-CONV on top of the A-RNN.

generated in the output response), where darker shades indicate larger attention weights. As the figure indicates, the attention mechanism helps learn a better RNN language model that promotes topic coherence, by learning to associate the currently-generated word with informative context words in the conversation history. As shown in Figure 4a, the A-RNN generates meaningful and topically coherent responses on the *MovieTriples* dataset. In comparison, the vanilla RNN tends to produce generic answers, such as “i don’t know”. Similarly, the A-RNN follows up with useful questions on the *Ubuntu Troubleshoot* dataset (Fig. 4b).

Conclusion

We investigate how to improve the performance of a recurrent neural network dialogue model via an attention mechanism, and how to promote topic and saliency aware conversation continuation. Our attention-RNN language model increases the scope of attention continuously as the conversation progresses (which distinguishes it from standard attention with fixed scope in a sequence-to-sequence models) such that each generated word can be associated with its most related words in the conversation history. We evaluate this simple model on two large dialogue datasets (*MovieTriples* and *Ubuntu Troubleshoot*), and achieve the best results reported to-date on multiple dialogue metrics (including complementary diversity-based metrics), performing better than gate-based RNN memory models. We also promote topic concentration by adopting LDA-based reranking, further improving performance.

Acknowledgments

We thank Iulian Serban, Yi Luan, and the anonymous reviewers for sharing their datasets and for their helpful discussion. We thank NVIDIA Corporation for donating GPUs used in this research.

Appendix

LDA-based Re-Ranking Details

While the trained attention-RNN dialogue model generates a natural language continuation of a conversation while maintaining topic concentration by token saliency/association, some dialogue-level topic-supervision can help to encourage generations that are more topic-aware. Such supervision is not commonly available, and we use unsupervised methods to learn document-level latent topics, and adopt the learned topic model to select the best continuation based on document-level topic-matching.

We choose Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003; Blei and Lafferty 2009) due to its demonstrated ability to learn a distribution over latent topics given a collection of documents. This generative model assumes that documents arise from multiple topics, each of which is defined as a distribution over a fixed vocabulary of terms. Specifically, we can assume that K topics are associated with a collection of dialogues $\{w_{0:T_n}\}_{n=1}^N$ in the dialogue corpus, and that each dialogue exhibits these topics with different likelihoods $\theta_n \in \mathbb{R}^K$. Let $\text{Dir}_V(\eta)$ and $\text{Dir}_K(\xi)$ denote a V - and K -dimensional Dirichlet distributions with

scalar parameter η and vector parameter $\xi \in \mathbb{R}^K$, respectively. The associated η and ξ can be learned by various methods (Blei, Ng, and Jordan 2003; Blei and Lafferty 2009; Hoffman, Bach, and Blei 2010), and the topic proportions $\hat{\theta}_n$ of a (possibly unseen) dialogue $w_{0:T_n}$ can be estimated with the learned topic structure \mathcal{L} . We let $\hat{\theta} = \mathcal{L}(w_{0:T})$ denote the estimated topic proportions $\hat{\theta}$ of any dialogue $w_{0:T}$ using the learned topic structure \mathcal{L} .

We use the training set to learn the topic structure \mathcal{L} with K topics. Having generated a set of continuations $\{c_m\}_{m=1}^M$ for each unseen dialogue $w_{0:T}$ during testing, we use the topic structure \mathcal{L} to compute the topic proportions for the dialogue and its continuations, $\hat{\theta}(w_{0:T}) = \mathcal{L}(w_{0:T})$ and $\hat{\theta}(c_m) = \mathcal{L}(c_m)$, respectively. Next, we employ a pre-defined metric $\mathcal{S}_m = \mathcal{S}(\hat{\theta}(w_{0:T}), \hat{\theta}(c_m))$ to compute the similarity between the topic proportions for the dialogue $\hat{\theta}(w_{0:T})$ and each continuation $\hat{\theta}(c_m)$. Finally, we compute a weighted score $\bar{\mathcal{S}}_m = \lambda \mathcal{S}_m + (1 - \lambda) \ell(c_m | w_{0:T})$, where $\lambda \in [0, 1]$ and $\ell(c_m | w_{0:T})$ is the log-likelihood of the continuation c_m conditioned on the dialogue $w_{0:T}$. The hyperparameters K and λ are tuned on a held-out development set according to the target evaluation metrics (defined shortly).

Concurrent with our work, Luan, Ji, and Ostendorf (2016) also propose using learned topic representations $\hat{\theta}$ of the given conversation as an extra feature in a language model in order to enhance the topic coherence of the generation. They define the function g as $g(h_t, \hat{\theta}, v_j) = O_{v_j}^\top (O_h h_t + O_\theta \hat{\theta})$. As we show in the Results and Analysis section, our attention-RNN language models significantly outperform this approach, though it is possible to further improve the quality of the generation by combining these methods i.e., by defining $g(h_t, z_t, \hat{\theta}, v_j) = O_{v_j}^\top (O_h h_t + O_z z_t + O_\theta \hat{\theta})$ and also using an LDA-based re-ranker.

Dataset Details

We train and evaluate the models on two large natural language dialogue datasets, *MovieTriples* (pre-processed by Serban et al. (2016)) and *Ubuntu Troubleshoot* (pre-processed by Luan, Ji, and Ostendorf (2016)). The dialogue within each of these datasets consists of a sequence of utterances (turns), each of which is a sequence of tokens (words). Table 7 provides the statistics for these two datasets.⁵

Evaluation Metrics

For the sake of comparison, we closely follow previous work and adopt several standard (and complementary) evaluation metrics (perplexity, word error rate, recall@N and BLEU, diversity-based Distinct-1). On the *MovieTriples* dataset, we use perplexity (PPL) and word error rate (WER) as is done in previous work, due to the metrics’ ability to appropriately quantify the performance of a language model by measuring

⁵Following the approach of Luan, Ji, and Ostendorf (2016), we randomly sample nine utterances as negative examples of the last utterance for each conversation in the *Ubuntu Troubleshoot* for the development set.

Table 7: Dataset Statistics

	<i>MovieTriples</i>	<i>Ubuntu</i>
training size	196308	216129
dev size	24717	13522
test size	24271	10020
turns / dialogue	3	6–20
avg. tokens / dialogue	53	116
vocabulary size	10003	20003

its ability to learn the syntactic structure of dialogue. Following Serban et al. (2016), we adopt two versions for each metric: i) PPL as the word-level perplexity over the entire dialogue conversation; ii) PPL@L as the word-level perplexity over the last utterance of the conversation; iii) WER as the word error rate over the entire dialogue conversation; and iv) WER@L as the word error rate over the last utterance of the conversation. We use all four metrics for model selection on the development set and for evaluation on the test set.

On the *Ubuntu Troubleshoot* dataset, we follow previous work and use PPL and recall@N to evaluate the performance of the proposed language models. recall@N (Manning et al. 2008) evaluates a model by measuring how often the model ranks the correct dialogue continuation within top-N range given 10 candidates. Additionally, we also adopt BLEU score (Papineni et al. 2001), to evaluate the quality of generations produced by the models. While not perfect, we believe that BLEU score is a suitable metric for evaluating topic-concentration, as it evaluates the generated continuation by counting explicit n-gram match relative to a set of references, and because word-matching can be a reasonable proximity for topic-matching.

Finally, we also present evaluation using the diversity-based Distinct-1 metric (proposed by Li et al. (2016a)) to measure the ability of the A-RNN to promote diversity in the generations, because typical neural dialogue models generate generic, safe responses (technically appropriate but not informative, e.g., “I don’t know”). Distinct-1 is computed as the number of distinct unigrams in the generation scaled by the total number of generated tokens.

Training Details

For the *MovieTriples* dataset, we follow the same procedure as Serban et al. (2016) and first pretrain on a large *Q-A SubTitle* dataset (Ameixa et al. 2014) and then fine-tune on the target *MovieTriples* dataset. The *Q-A SubTitle* dataset contains approximately 5.5M question-answer pairs, from which we randomly sample 20000 pairs as the held-out set. We performed early-stopping according to the PPL score on the held-out set. We tuned the dimensionality $d \in \{200, 300, 400, 500\}$ of the RNN hidden states, and chose $d = 300$. We similarly use a setting of $d = 300$ on the *Ubuntu Troubleshoot* dataset. We use Adam (Kingma and Ba 2015) for optimization in RNN back-propagation.

When training the attention-RNN dialogue model, we perform early-stopping and model selection based on the PPL values on the development set, following Luan, Ji, and Ostendorf (2016). For the LDA-based re-ranker, we choose

History:	A: i knew how much you wanted it . R: but i was going to buy it for myself .
A-RNN:	A: i ’ d like to buy you a drink .
RNN:	A: i don ’ t know .
Reference:	A: i wanted to give it to you . for our anniversary . we met five years ago today .
History:	A: person , but this time i got the money . R: if you give it to me , i might forget i found you .
A-RNN:	A: i ’ ll give it to you .
RNN:	A: i don ’ t know .
Reference:	A: i don ’ t have it with me . tell person .
(a) <i>MovieTriples</i>	
History:	A: so i need to make an ntfs partition help ? R: install ntfsprogs R: it will add the functionality to gparted
A-RNN:	R: sudo apt-get install ntfsprogs
RNN:	R: how do i do that ?
Reference:	R: sudo apt-get update ; sudo apt-get install ntfsprogs
History:	A: i have really choppy streaming video in 9.10 , any way to fix that ? A: any way to fix choppy streaming video in 9.10 ? R: what kind of video
A-RNN:	R: what video card do you have ?
RNN:	R: what are you trying to do ?
Reference:	R: what cpu do you have
(b) <i>Ubuntu Troubleshoot</i>	

Figure 4: Examples of responses generated by the A-RNN along with the reference response.

the two hyper-parameters, i.e., the number of topics K and ranking weight λ , based on the values of recall@N and BLEU score on the development set. In the end, we choose $K = 10$ out of $K \in \{5, 10, 20, 50\}$ and $\lambda = 0.45$ out of $\lambda \in \{0.00, 0.05, \dots, 0.95, 1.00\}$.

Output Examples

Next, we qualitatively evaluate the effectiveness of our A-RNN model through visualizations of the output on both datasets. As is evident in Figure 4a, the A-RNN generates meaningful and topically coherent responses on the *MovieTriples* dataset. In comparison, the vanilla RNN tends to produce generic answers, such as “i don’t know”. Similarly, the A-RNN either provides promising technical solutions or follows up with useful questions on the *Ubuntu Troubleshoot* dataset (Fig. 4b).

References

- Ameixa, D.; Coheur, L.; Fialho, P.; and Quaresma, P. 2014. Luke, I am your father: dealing with out-of-domain requests by using movie subtitles. In *Intelligent Virtual Agents*, 13–21. Springer.
- Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*.
- Blei, D. M., and Lafferty, J. D. 2009. Topic models. *Text mining: Classification, clustering, and applications* 10(71):34.
- Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *the Journal of machine Learning research*.
- Crow, B. 1983. Topic shifts in couples' conversations. *Conversational coherence: Form, structure, and strategy* 136–156.
- Dušek, O., and Jurčiček, F. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of SIGDIAL*.
- Goldberg, J. 1983. A move towards describing conversational coherence. *Conversational coherence: Form, structure, and strategy* 25–45.
- Hoffman, M.; Bach, F. R.; and Blei, D. M. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*.
- Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016b. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.
- Luan, Y.; Ji, Y.; and Ostendorf, M. 2016. LSTM based conversation models. *arXiv preprint arXiv:1603.09457*.
- Manning, C. D.; Raghavan, P.; Schütze, H.; et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.
- Mei, H.; Bansal, M.; and Walter, M. R. 2016a. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- Mei, H.; Bansal, M.; and Walter, M. R. 2016b. What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT)*.
- Mikolov, T. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*.
- Mnih, V.; Hees, N.; Graves, A.; and Kavukcuoglu, K. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems (NIPS)*.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.
- Parikh, A. P.; Täckström, O.; Das, D.; and Uszkoreit, J. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.
- Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Sanders, R. 1983. Tools for cohering discourse and their strategic utilization. *Conversational coherence: Form, structure, and strategy* 67–80.
- Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural networks. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.
- Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Sigman, S. 1983. Some multiple constraints placed on conversational topics. *Conversational coherence: Form, structure, and strategy*.
- Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Sutskever, I.; Vinyals, O.; and Lee, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.
- Tran, K.; Bisazza, A.; and Monz, C. 2016. Recurrent memory network for language modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT)*.
- Vinyals, O., and Le, Q. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.
- Wen, T.-H.; Gašić, M.; Mrkšić, N.; Su, P.-H.; Vandyke, D.; and Young, S. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Wen, T.-H.; Gašić, M.; Mrkšić, N.; M. Rojas-Barahona, L.; Su, P.-H.; Vandyke, D.; and Young, S. 2016. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of the Conference on North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics.
- Winograd, T. 1971. Procedures as a representation for data in a computer program for understanding natural language. Technical report, DTIC Document.
- Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Yao, K.; Peng, B.; Zweig, G.; and Wong, K.-F. 2016. An attentional neural conversation model with improved specificity. *arXiv preprint arXiv:1606.01292*.