Goal-Embedded Dual Hierarchical Model for Task-Oriented Dialogue Generation

Yi-An Lai
Amazon ML
Seattle
yianl@amazon.com

Arshit Gupta
Amazon ML
Seattle
arshig@amazon.com

Yi Zhang
Amazon ML
Seattle
yizhngn@amazon.com

Abstract

Hierarchical neural networks are often used to model inherent structures within dialogues. For goal-oriented dialogues, these models miss a mechanism adhering to the goals and neglect the distinct conversational patterns between two interlocutors. In this work, we propose *Goal-Embedded Dual Hierarchical Attentional Encoder-Decoder (G-DuHA)* able to center around goals and capture interlocutor-level disparity while modeling goal-oriented dialogues. Experiments on dialogue generation, response generation, and human evaluations demonstrate that the proposed model successfully generates higher-quality, more diverse and goal-centric dialogues. Moreover, we apply data augmentation via goal-oriented dialogue generation for task-oriented dialog systems with better performance achieved.

1 Introduction

Modeling a probability distribution over word sequences is a core topic in natural language processing, with language modeling being a flagship problem and mostly tackled via recurrent neural networks (RNNs) (Mikolov and Zweig, 2012; Melis et al., 2017; Merity et al., 2018).

Recently, dialogue modeling has drawn much attention with applications to response generation (Serban et al., 2016a; Li et al., 2016b; Asghar et al., 2018) or data augmentation (Yoo et al., 2019). It's inherently different from language modeling as the conversation is conducted in a turn-by-turn nature. (Serban et al., 2016b) imposes a hierarchical structure on encoder-decoder to model this utterance-level and dialogue-level structures, followed by (Serban et al., 2016c; Chen et al., 2018; Le et al., 2018a).

However, when modeling dialogues involving two interlocutors center around one or more goals, these systems generate utterances with the greatest likelihood but without a mechanism sticking to the goals. This makes them go off the rails and fail to model context-switching of goals. Most of the generated conversations become off-goal dialogues with utterances being non-relevant or contradicted to goals rather than on-goal dialogues. The differences are illustrated in Figure 1.

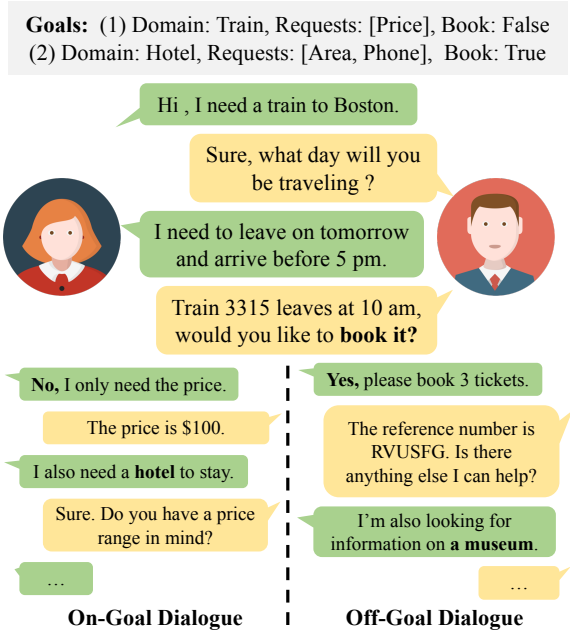

Figure 1: On-goal dialogues follow the given goals such as no booking of train tickets and a hotel reservation. Off-goal dialogues have context switches to other domains non-relevant to goals.

Besides, two interlocutors in a goal-oriented dialogue often play distinct roles as one has requests or goals to achieve and the other provides necessary support. Modeled by a single hierarchical RNN, this interlocutor-level disparity is neglected and constant context switching of roles could reduce the capacity for tracking conversational flow and long-term temporal structure.

To resolve the aforementioned issues when modeling goal-oriented dialogues, we propose

the Goal-Embedded Dual Hierarchical Attentional Encoder-Decoder (G-DuHA) to tackle the problems via three key features. First, the goal embedding module summarizes one or more goals of the current dialogue as goal contexts for the model to focus on across a conversation. Second, the dual hierarchical encoder-decoders can naturally capture interlocutor-level disparity and represent interactions of two interlocutors. Finally, attentions are introduced on word and dialogue levels to learn temporal dependencies more easily.

In this work, our contributions are that we propose a model called goal-embedded dual hierarchical attentional encoder-decoder (G-DuHA) to be the first model able to focus on goals and capture interlocutor-level disparity while modeling goal-oriented dialogues. With experiments on dialogue generation, response generation and human evaluations, we demonstrate that our model can generate higher-quality, more diverse and goal-focused dialogues. In addition, we leverage goal-oriented dialogue generation as data augmentation for task-oriented dialogue systems, with better performance achieved.

2 Related Work

Dialogues are sequences of utterances, which are sequences of words. For modeling or generating dialogues, hierarchical architectures are usually used to capture their conversational nature. Traditionally, language models are also used for modeling and generating word sequences. As goal-oriented dialogues are generated, they can be used in data augmentation for task-oriented dialogue systems. We review related works in these fields.

Dialogue Modeling. To model conversational context and turn-by-turn structure of dialogues, (Serban et al., 2016b) devised hierarchical recurrent encoder-decoder (HRED). Reinforcement and adversarial learning are then adopted to improve naturalness and diversity (Li et al., 2016b, 2017a). Integrating HRED with the latent variable models such as variational autoencoder (VAE) (Kingma and Welling, 2014) extends another line of advancements (Serban et al., 2016c; Zhao et al., 2017; Park et al., 2018; Le et al., 2018b). However, these systems are not designed for task-oriented dialogue modeling as goal information is not considered. Besides, conversations between two interlocutors are captured with a single encoder-decoder by these systems.

Language Modeling. A probability distribution of a word sequence $w_{1:T} = (w_1, w_2, ..., w_T)$ can be factorized as $p(w_1)\prod_{t=2}^{T} p(w_t|w_{1:t-1})$. To approximate the conditional probability $p(w_t|w_{1:t-1})$, counted statistics and smoothed N-gram models have been used before (Goodman, 2001; Katz, 1987; Kneser and Ney, 1995). Recently, RNN-based models have achieved a better performance (Mikolov et al., 2010; Józefowicz et al., 2016; Grave et al., 2017; Melis et al., 2018). As conversational nature is not explicitly modeled, models often have role-switching issues.

Task-Oriented Dialogue Systems. Conventional task-oriented dialog systems entails a sophisticated pipeline (Raux et al., 2005; Young et al., 2013) with components including spoken language understanding (Chen et al., 2016; Mesnil et al., 2015; Gupta et al., 2019), dialog state tracking (Henderson et al., 2014; Mrksic et al., 2017), and dialog policy learning (Su et al., 2016; Gašić and Young, 2014). Building a task-oriented dialogue agent via end-to-end approaches has been explored recently (Li et al., 2017b; Wen et al., 2017). Although several conversational datasets are published recently (Gopalakrishnan et al., 2019; Henderson et al., 2019), the scarcity of annotated conversational data remains a key problem when developing a dialog system. This motivates us to model task-oriented dialogues with goal information in order to achieve controlled dialogue generation for data augmentation.

3 Model Architecture

Given a set of goals and the seed user utterance, we want to generate a goal-centric or on-goal dialogue that follows the domain contexts and corresponding requests specified in goals. In this section, we start with the mathematical formulation, then introduce our proposed model, and describe our model's training objective and inference.

At training time, K dialogues $\{D_1, ..., D_K\}$ are given where each D_i associates with N_i goals $\mathbf{g_i} = \{g_{i1}, g_{i2}, ..., g_{iN_i}\}$. A dialogue D_i consists of M turns of utterances between a user $\mathbf{u}$ and a system agent $\mathbf{s}$ $(\mathbf{w_{u1}}, \mathbf{w_{s1}}, \mathbf{w_{u2}}, \mathbf{w_{s2}}, ...)$, where $\mathbf{w_{u1}}$ is a word sequence $w_{u1,1}, w_{u1,2}, ..., w_{u1,N_{u1}}$ denoting the user's first utterance.

The task-oriented dialogue modeling aims to approximate the conditional probability of user's or agent's next utterance given previous turns and goals. It can be further decomposed over gener-

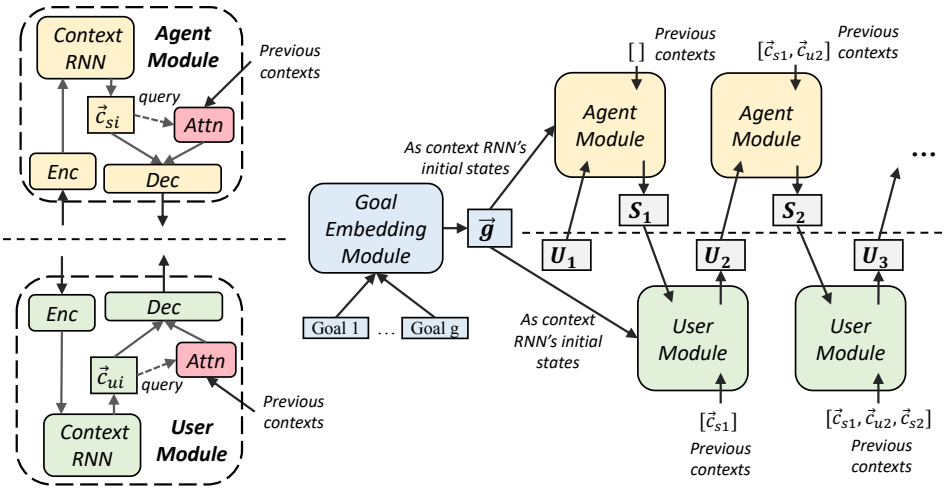

Figure 2: G-DuHA architecture. The goal embedding module embeds goals as priors for context RNNs. Dual hierarchical RNNs naturally model two interlocutors. An attention over previous contexts captures long-term dependencies. For encoders, word attentions are used to summarize local importance of words. (Enc: Encoder, Dec: Decoder, Attn: Attention, U_i: User's utterance, S_i: Agent's utterance)

ated words, e.g.

$$P(\mathbf{w_{um}}|\mathbf{w_{u1}}, \mathbf{w_{s1}}, ..., \mathbf{w_{s(m-1)}}, \mathbf{g_i}) = \prod_{n=1}^{N_{um}} P(w_{um,n}|w_{um,<n}, \mathbf{w_{u1}}, ...) \quad (1)$$

To model goal-oriented dialogues between two interlocutors, we propose Goal-embedded Dual Hierarchical Attentional Encoder-Decoder (G-DuHA) as illustrated in Fig. 2. Our model comprises goal embedding module, dual hierarchical RNNs, and attention mechanisms detailed below.

3.1 Goal Embedding Module

We represent each goal in $\{g_{i1}, g_{i2}, ...\}$ using a simple and straightforward binary or multi-one-hot encoding followed by a feed-forward network (FFN). A goal could have a specific domain such as hotel or a request such as price or area, Table 2 shows a few examples. Our goal embedding module, a FFN, then converts binary encoding of each goal into a goal embedding, where the FFN is learned during training. If multiple goals present, all goal embeddings are then added up element-wisely to be the final goal embedding:

$$\overrightarrow{g_i} = \sum_{j=1}^{|g_i|} FFN(Encode(g_{ij})) \quad (2)$$

The output of the goal embedding module has the same number of dimensions as context RNN's hidden state and is used as initial states for all layers of all context RNNs to inform the model about a set of goals to focus on.

3.2 Dual Hierarchical Architecture

The hierarchical encoder-decoder structure (Serban et al., 2016b) is designed for utterance level and context level modeling. With a single encoder, context RNN, and decoder, the same module is used to process input utterances, track contexts, and generate responses for both interlocutors.

In task-oriented dialogues, however, roles are distinct as the user aims to request information or make reservations to achieve goals in mind and the system agent provides necessary help. To model this interlocutor-level disparity, we extend it into a dual architecture involving two hierarchical RNNs, each serves as a role in a dialogue.

3.3 Attention Mechanisms

At the utterance level, as importance of words can be context dependent, our model uses first hidden layer's states of the context RNN as the query for attention mechanisms (Bahdanau et al., 2015; Xu et al., 2015) to build an utterance representation. A feed-forward network is involved to computed attention scores, whose input is the concatenation of the query and an encoder output.

At the dialogue level, for faster training, our model has a skip connection to add up context RNN's raw output with its input as the final output c_t. To model long-term dependencies, an attention module is applied to summarize all previous

contexts into a global context vector. Specifically, a feed-forward network takes the current context RNN's output c_t as the query and all previous context outputs from both context RNNs as keys and values to compute attention scores. The global context vector is then concatenated with c_t to form our final context vector for decoder to consume.

3.4 Objective

For predicting the end of a dialogue, we exploit a feed-forward network over the final context vector for a binary prediction. Thus our training objective can be written as

$$L = \sum_{i=1}^{K} \Big[-\log p_{\boldsymbol{\theta}}(D_i) + \sum_{t}^{M_i} -\log p_{\boldsymbol{\theta}}^{end}(e_{it}) \Big], \tag{3}$$

where our model $p_{\boldsymbol{\theta}}$ has parameters $\boldsymbol{\theta}$, M_i is the number of turns, e_{it} is 0 as the dialogue D_i continues and 1 if it terminates at turn t.

3.5 Generation

At dialogue generation time, a set of goals $\{g_{i1}, g_{i2}, ...\}$ and a user utterance $\mathbf{w_{u1}}$ as a seed are given. Then our model will generate conversations simulating interactions between a user and an agent that seek to complete all given goals. The generation process terminates as the end of dialogue prediction outputs a positive or the maximum number of turns is reached.

4 Experiments

We evaluate our approach on dialogue generation and response generation as well as by humans. Ablation studies and an extrinsic evaluation that leverages dialogue generation as a data augmentation method are reported in the subsequent section.

4.1 Dataset

Experiments are conducted on a task-oriented human-human written conversation dataset called MultiWOZ (Budzianowski et al., 2018), the largest publicly available dataset in the field. Dialogues in the dataset span over diverse topics, one to more goals, and multiple domains such as `restaurant`, `hotel`, `train`, etc. It consists of 8423 train dialogues, 1000 validation and 1000 test dialogues with on average 15 turns per dialogue and 14 tokens per turn.

4.2 Baselines

We compare our approach against four baselines:

(i) LM+G: As long-established methods for language generation, we adopt an RNN language model (LM) with 3-layer 200-hidden-unit GRU (Cho et al., 2014) incorporating our goal embedding module as a baseline, which has goal information but no explicit architecture for dialogues.

(ii) LM+G-XL: To show the possible impact of model size, a larger LM that has a 3-layer 450-hidden-unit GRU is adopted as another baseline.

(iii) Hierarchical recurrent encoder-decoder (HRED) (Serban et al., 2016b): As the prominent model for dialogues, we use HRED as the baseline that has a dialogue-specific architecture but no goal information. The encoder, decoder, and context RNN are 2-layer 200-hidden-unit GRUs.

(iv) HRED-XL: We also use a larger HRED with 350 hidden units for all GRUs as a baseline to show the impact of model size.

4.3 Implementation Details

In all experiments, we adopt the delexicalized form of dialogues as shown in Table 2 with vocabulary size, including slots and special tokens, to be 4258. The max number of turns and sequence length are capped to 22 and 36, respectively.

G-DuHA uses 2-layer, 200-hidden-unit GRUs as all encoders, decoders, and context RNNs. All feed-forward networks have 2 layers with non-linearity. FFNs of encoder attention and end of dialogue prediction have 50 hidden units. The FFNs of context attention and goal embedding gets 100 and 200 hidden units. We simply use greedy decoding for utterance generation.

All models initialize embeddings with pre-trained fast-text vectors on wiki-news (2018) and are trained by the Adam optimizer (2015) with early-stopping to prevent overfitting. To mitigate the discrepancy between training and inference, we pick predicted or ground-truth utterance as the current input uniformly at random when training.

4.4 Evaluation Metrics

We employ a number of automatic metrics as well as human evaluations to benchmark competing models on quality, diversity, and goal focus:

Quality. BLEU (Papineni et al., 2002), as BLEU-4 by default, is a word-overlap measure against references and commonly used by dialogue generation works to evaluate quality (2015;

Model	Size	BLEU	B1	B2	B3	D-1	D-2	D-U	P	R	F1	L-D	L-U
LM+G	4.2 M	6.34	23.24	12.89	8.76	0.16	0.88	23.75	89.52	82.39	**84.89**	15.1	13.0
LM+G-XL	8.2 M	6.22	23.10	12.75	8.63	0.16	0.93	**26.38**	**90.03**	81.62	84.71	14.7	14.4
HRED	5.1 M	5.40	21.91	11.58	7.66	0.09	0.38	3.94	69.69	66.22	65.35	17.3	14.1
HRED-XL	8.8 M	5.08	20.45	10.90	7.22	0.11	0.50	5.51	68.25	65.93	65.09	15.9	16.9
G-DuHA	7.8 M	**7.48**	**24.38**	**14.44**	**10.12**	**0.17**	**0.99**	**25.98**	86.21	**83.18**	83.76	14.3	15.0

Table 1: **Dialogue** generation performance. All results are averaged over 5 random seeds and reported in percentage (%). **Size**: Total number of model parameters, **P**: Precision, **R**: Recall, **L-D**: Average dialogue length, **L-U**: Average utterance length.

Goals	LM + Goal	HRED	G-DuHA
Domain: attraction User: [name] Reqt: [**entrance fee**] Book: False Domain: restaurant User: [area, food] Reqt: [] **Book: True** Domain: **taxi** User: [arriveBy] Reqt: [car type, **phone**] Book: False	U: what is the address for [attraction_name] ? S: sure ! [attraction_name] is located at [attraction_address] and **is free .** U: i . i am also **looking for a place to eat .** the [value_area] . S: there are [value_count] restaurant -s in the [value_area] . what type of food are you looking for ? ... U: yes , **please book a table** for [value_count] people at [value_time] on [value_day] . ... **U: i help you with anything else ?** S: i **also need a taxi** to get me the restaurant to the restaurant . U: i time would you like to leave ? S: i would like to arrive by the restaurant by the booked time . U: booking completed ! booked car type [taxi_type] **contact number [taxi_phone]** S: thank you that s all i needed . U: you are welcome . have a great day !	U: what is the address for [attraction_name] ? S: the address is [hospital_address] , [value_place] . is there anything else i can help you with ? U: yes , i would like to book a table for [value_count] at [value_time] on [value_day] . S: i **have booked your table .** your reference number is [restaurant_reference] . U: thank you . i also need **a train** to [value_place] that leaves after [value_time] . S: there are [value_count] trains that match your request . what day and time would you like to travel ? U: i would like to leave on [value_day] and arrive by [value_time] . S: there are [value_count] trains that meet your criteria . would you like to narrow it down by departure time or arrival time ? U: i would like to arrive by [value_time]	U: what is the address for [attraction_name] ? S: the address is [attraction_address] and the postcode is [attraction_postcode] . U: **what is the entrance fee ?** S: the **entrance fee is free** . is there anything else i can help you with ? U: yes , i am also **looking for a restaurant** that serves [value_food] food in the [value_area] . ... S: i have [value_count] restaurant -s that meet your criteria . would you like me to **book you a table ?** **U: yes , please .** i need a table for [value_count] people at [value_time] on [value_day] . ... U: i **also need a taxi** to commute between the [value_count] place -s . S: i have booked you a taxi to pick you up at the [attraction_name] . it will be a [taxi_type] and the **contact number is [taxi_phone]** . U: thank you , that is all i need . S: you are welcome . have a great day !

Table 2: Dialogue qualitative comparison. Reqt: Requests. U: User, S: Agent. **Goal hit or miss**. **Role confusion**. Extensive qualitative comparisons of dialogues are presented in the appendix.

2016b; 2016a; 2017; 2018). Lower N-gram B1, B2, B3 are also reported.

Diversity. D-1, D-2, D-U: The distinctiveness denotes the number of unique unigrams, bigrams, and utterances normalized by each total count (Li et al., 2016a; Xu et al., 2018). These metrics are commonly used to evaluate the dialogue diversity.

Goal Focus. A set of slots such as address are extracted from reference dialogues as multi-label targets. Generated slots in model's output dialogues are the predictions. We use the multi-label precision, recall, and F1-score as surrogates to measure the goal focus and achievement.

Human Evaluation. The side-by-side human preference study evaluates dialogues on *goal focus*, *grammar*, *natural flow*, and *non-redundancy*.

5 Results and Discussion

5.1 Dialogue Generation Results

For dialogue generation (Li et al., 2016b), a model is given one or more goals and one user utterance as the seed inputs to generate entire dialogues in an auto-regressive manner.

Table 1 summarizes the evaluation results. For quality measures, G-DuHA significantly outperforms other baselines, implying that it's able to carry out a higher-quality dialogue. Besides, goal-embedded LMs perform better than HREDs, showing the benefits of our goal embedding module. No significant performance difference is observed with respect to model size variants.

For diversity evaluations, G-DuHA is on par with goal-embedded LMs and both outperform HRED significantly. Of 1000 generated dialogues, HRED delivers highly repetitive outputs with only 4 to 6% distinct utterances, whereas 25% of utterances are unique from G-DuHA.

For recovering slots in reference dialogues, precision denotes a degree of goal deviation, recall entails the achievement of goals, and F1 measures the overall focus. Goal-embedded LM is the best on precision and F1 with G-DuHA having com-

Model	BLEU	B1	B2	B3	D-1	D-2	D-U	P	R	F1	L-R
LM+G	14.88	35.86	24.59	18.81	0.27	1.44	40.84	79.71	68.57	71.73	14.3
LM+G-XL	14.51	35.28	24.07	18.36	**0.28**	**1.47**	**42.56**	**79.79**	67.31	71.00	14.3
HRED	14.34	36.27	24.31	18.33	0.21	0.94	20.21	75.46	67.08	68.78	17.1
HRED-XL	14.33	36.36	24.37	18.35	0.23	1.12	26.63	72.69	68.24	68.20	17.3
G-DuHA	**15.85**	**37.99**	**26.14**	**20.01**	0.25	1.27	**39.59**	**78.34**	**71.55**	**72.69**	16.7

Table 3: **Agent's response** generation performance. All results are averaged over 5 random seeds and reported in percentage (%). **P**: Precision, **R**: Recall, **L-R**: Average response length.

Model	BLEU	B1	B2	B3	D-1	D-2	D-U	P	R	F1	L-R
LM+G	11.73	31.79	21.26	15.56	0.35	1.82	33.57	89.44	75.78	80.23	10.6
LM+G-XL	11.60	31.49	21.00	15.38	**0.36**	**1.87**	34.29	89.57	75.55	80.03	10.7
HRED	10.88	31.69	20.46	14.65	0.24	0.98	16.00	80.00	79.11	77.58	13.1
HRED-XL	10.81	31.84	20.48	14.60	0.26	1.15	19.87	80.11	78.82	77.42	13.2
G-DuHA	**13.25**	**35.20**	**23.89**	**17.56**	0.30	1.49	**35.57**	**91.12**	**79.66**	**83.51**	12.8

Table 4: **User's response** generation performance. All results are averaged over 5 random seeds and reported in percentage (%). **P**: Precision, **R**: Recall, **L-R**: Average response length.

Goals	Contexts (Reference)	Response
Domain: train User: [destination, departure, arriveBy, day] Reqt:[duration,price] Book: False Domain: restaurant User: [area, food, pricerange] Reqt: [] Book: True	U: hey i am looking for a train from [value_place] to [value_place] . mind helping out ? S: many trains are leaving on [value_day] . mind choosing the time ? U: i want to leave on [value_day] and arrive by [value_time] . S: i have train [train_id] that leaves [value_place] at [value_time] and arrives in [value_place] at [value_time] . would you like to make reservations ? U: could you give me the travel time and price of that train please ? S: the cost is [value_price] pounds . the duration of the trip is [value_count] minutes , leaving [value_place] at [value_time] and arriving at [value_time] in [value_place] .	**LM+G**: thanks . i am also looking for a restaurant in the [value_area] . **HRED**: i am also looking for a place to stay . i would like a [value_count] star hotel . **G-DuHA**: thank you . i am also looking for a restaurant in the [value_area] of town .
Domain: train User: [destination, departure, arriveBy, day] Reqt: [] Book: False	U: i need a train that leaves on [value_day] with an arrival of [value_time] . S: where will you be travelling to and from ?	**LM+G**: i will be departing from [value_place] . **HRED**: i will be leaving from [value_place] . **G-DuHA**: i am departing from [value_place] and need to arrive by [value_time] .

Table 5: Qualitative comparison of generated responses. Reqt: Requests. U: User, S: Agent.

parable performance. However, even thought LM can better mention the slots in dialogue generation, utterances are often associated with a wrong role. That is, role confusions are commonly seen such as the user makes reservations for the agent as in Table 2. The reason could be that LM handles the task similar to paragraph generation without an explicit design for the conversational hierarchy.

Overall, G-DuHA is able to generate high-quality dialogues with sufficient diversity and still adhere to goals compared to baselines.

Qualitative Comparison. Table 2 compares generated dialogues from different models given one to three goals to focus on. It's clear that models with the goal embedding module are able to adhere to given goals such as "book" or "no book", requesting "price" or "entrance fee" while HRED fails to do so. They can also correctly covering all required domain contexts without any diversion such as switching from attraction inquiry to restaurant booking, then to taxi-calling. For HRED, without goals, generated dialogues often detour to a non-relevant domain context such as shifting to train booking while only hotel inquiry is required.

For goal-embedded LM, a serious issue revealed is role confusions as LM often wrongly shifts between the user and agent as shown in Table 2. The issue results from one wrong `EndofUtterance` prediction but affects rest of dialogue and degrades the overall quality. More generated dialogues are reported in the appendix.

5.2 Response Generation Results

For response generation (Sordoni et al., 2015; Serban et al., 2016c; Park et al., 2018), a set of goals as well as the previous context, i.e. all previous

	Wins	Losses	Ties
Goal Focus	**82.33%**	6.00%	11.67%
Grammar	6.00%	5.00%	89.00%
Natural Flow	**26.00%**	15.00%	59.00%
Non-redundancy	**35.34%**	6.33%	58.33%

Table 6: Human evaluations, G-DuHA vs HRED. 100 pairs of generated dialogues along with goals are given to three domain experts for side-by-side comparisons.

reference utterances, are given to a model to generate the next utterance as a response.

Table 3 and 4 summarize the results. G-DuHA outperforms others on quality and goal focus measures and rivals LM-goal on diversity on both agent and user responses. For goal focus, LM-goal performs good on precision but short on recall. This could because it generates much shorter user and agent responses on average.

Interestingly, as previous contexts are given, LM-goal performs only slightly better than HRED. This implies hierarchical structures capturing longer dependencies can make up the disadvantages of having no goal information for response generation. However, as illustrated in Table 5, HRED could still fail to predict the switch of domain contexts, e.g. from `train` to `restaurant`, which explains performance gaps. Another intriguing observation is that when incorporating the goal embedding module, response diversity and goal focus can be boosted significantly.

Comparing the performance between agent and user response generation, we observe that models can achieve higher quality and diversity but lower goal focus when modeling agent's responses. These might result from the relatively consistent utterance patterns but diverse slot types used by an agent. More generated responses across different models are presented in the appendix.

5.3 Human Evaluation Results

For human evaluation, we conduct side-by-side comparisons between G-DuHA and HRED, the widely used baseline in literature, on dialogue generation task. We consider the following four criteria: *goal focus*, *grammar*, *natural flow*, and *non-redundancy*. *Goal focus* evaluates whether the dialogue is closely related to the preset goals; *grammar* evaluates whether the utterances are well-formed and understandable; *natural flow* evaluates whether the flow of dialogue is logical and fluent; and *non-redundancy* evaluates whether the dialogue is absent of unnecessary repetition of mentioned information. 100 pairs of generated dialogues from G-DuHA and HRED along with their goals are randomly placed against each other. For each goal and pair of dialogues, three domain experts were instructed to set their preferences with respect to each of the four criteria, marked as *win / lose / tie* between the dialogues.

Table 6 presents the results. G-DuHA shows substantial advantages on goal focus, with 82.33% wins over HRED, confirming the benefits of our goal embedding module. G-DuHA also outperforms HRED significantly on natural flow and non-redundancy. These might result from G-DuHA's ability to generating much more diverse utterances while concentrating on current goals. An especially interesting observation is that in cases where multiple goals are given, G-DuHA not only stays focused on each individual goal but also generates intuitive transitions between goals, so that the flow of a dialogue is natural and coherent. An example is shown in Table 2, where the G-DuHA-generated dialogue switches towards the `taxi` goal while maintaining reference to the previously mentioned `attraction` and `restaurant` goals: *"... i also need a taxi to commute between the 2 places ..."*. We also observe that both G-DuHA and HRED performed well on grammaticality. The generated samples across all RNN-based models are almost free from grammar error as well.

5.4 Ablation Studies

The ablation studies are reported in Table 7 for dialogue generation and in Table 8 for response generation to investigate the contribution of each module. Here we evaluate user and agent response generation together.

Goal Embedding Module. First, we examine the impact of goal embedding module. When unplugging the goal embedding module, we observe significant and consistent drops on quality, diversity, and goal focus measures for both dialogue and response generation tasks. For dialogue generation task, the drops are substantially large which resonates with our intuition as the model only has the first user utterance as input context to follow.

With no guideline about what to achieve and what conversation flow to go around with, dialogues generated from HRED often have the similar flow and low diversity. These results demon-

Model	BLEU	B1	D-1	D-2	D-U	P	R	F1
G-DuHA	7.48	24.38	0.17	0.99	25.98	86.21	83.18	83.76
w/o goal	5.19	20.04	0.13	0.68	13.83	69.18	68.21	66.86
w/o dual	7.34	24.99	0.15	0.79	19.22	85.24	82.62	82.96
w/o context attention	7.34	24.34	0.17	0.99	24.98	86.70	83.40	84.10

Table 7: Ablation studies on **dialogue** generation over goal embedding module, dual architecture, and dialogue-level attention. Results are averaged over 5 random seeds and reported in percentage (%). **P**: Precision, **R**: Recall.

Model	BLEU	B1	D-1	D-2	D-U	P	R	F1
G-DuHA	14.84	36.84	0.18	1.10	34.23	89.87	82.00	84.80
w/o goal	13.29	35.23	0.16	0.97	28.04	84.33	80.15	81.10
w/o dual	14.60	36.66	0.17	0.96	27.53	89.43	80.81	83.91
w/o context attention	14.73	36.88	0.18	1.14	34.55	90.28	81.54	84.80

Table 8: Ablation studies on **response** generation over goal embedding module, dual architecture, and dialogue-level attention. Results are averaged over 5 random seeds and reported in percentage (%). **P**: Precision, **R**: Recall.

	Joint Goal	Turn Request
GLAD	88.55%	97.11%
GLAD + LM+G	88.07%	96.02%
GLAD + HRED	89.03%	97.11%
GLAD + G-DuHA	**89.04%**	**97.59%***

Table 9: Test accuracy of GLAD (Zhong et al., 2018) on the WoZ restaurant reservation dataset with different data augmentation models. (*significant against others.)

strate that our goal embedding module is critical in generating higher-quality and goal-centric dialogues with much more diversity.

Dual Hierarchical Architecture. We also evaluate the impact of dual hierarchical architecture. Comparisons on both dialogue and response generation tasks show a consistent trend. We observe that applying dual architecture for interlocutor-level modeling leads to a solid increase in utterance diversity as well as moderate improvements on quality and goal focus.

The results echo our motivation as two interlocutors in a goal-oriented dialogue scenario exhibit distinct conversational patterns and this interlocutor-level disparity should be modeled by separate hierarchical encoder-decoders.

For the dialogue-level attention module, there is no significant effect on diversity and goal focus on both tasks but it marginally improves the overall utterance quality as BLEU scores go up by a bit.

6 Data Augmentation via Dialogue Generation

As an exemplified extrinsic evaluation, we leverage the goal-oriented dialogue generation as data augmentation for task-oriented dialogue systems. Dialogue state tracking (DST) is used as our evaluation task which is a critical component in task-oriented dialogue systems (Young et al., 2013) and has been studied extensively (Henderson et al., 2014; Mrksic et al., 2017; Zhong et al., 2018).

In DST, given the current utterance and dialogue history, a dialogue state tracker determines the state of the dialogue which comprises a set of *requests* and *joint goals*. For each user turn, the user informs the system a set of turn goals to fulfill, *e.g. inform(area=south)*, or turn requests asking for more information, *e.g. request(phone)*. The joint goal is the collection of all turn goals up to the current turn.

We use the state-of-the-art Global-Locally Self-Attentive Dialogue State Tracker (GLAD) (Zhong et al., 2018) as our benchmark model and the WoZ restaurant reservation dataset (Wen et al., 2017; Zhong et al., 2018) as our benchmark dataset, which is commonly used for the DST task.

The dataset consists of 600 train, 200 validation and 400 test dialogues. We use the first utterances from 300 train dialogues and sample restaurant-domain goals to generate dialogues, whose states are annotated by a rule-based method.

Table 9 summarizes the augmentation results. Augmentation with G-DuHA achieved an improvement over the vanilla dataset and outperform

HRED on turn requests while being comparable on joint goal. For goal-embedded LM, as it struggles with role confusion, the augmentation actually hurts the overall performance.

7 Conclusion

We introduced the goal-embedded dual hierarchical attentional encoder-decoder (G-DuHA) for goal-oriented dialogue generation. G-DuHA is able to generate higher-quality and goal-focused dialogues as well as responses with decent diversity and non-redundancy. Empirical results show that the goal embedding module plays a vital role in the performance improvement and the dual architecture can significantly enhance diversity.

We demonstrated one application of the goal-oriented dialogue generation through a data augmentation experiment, though the proposed model is applicable to other conversational AI tasks which remains to be investigated in the future.

As shown in experiments, a language model coupled with goal embedding suffers from role-switching or confusion. It's also interesting to further dive deep with visualizations (Kessler, 2017) and quantify the impact on quality, diversity, and goal focus metrics.

Acknowledgments

The authors would like to acknowledge the entire AWS Lex Science team for thoughtful discussions, honest feedback, and full support. We are also very grateful to the reviewers for insightful comments and helpful suggestions.

References

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *EMNLP*.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1653–1662. International World Wide Web Conferences Steering Committee.

Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Interspeech*, pages 3245–3249.

Kyunghyun Cho, Bart van Merrienboer, aglar Gülehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Milica Gašić and Steve Young. 2014. Gaussian processes for pomdp-based dialogue manager optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):28–40.

Joshua T Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tr. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2017. Improving neural language models with a continuous cache. *CoRR*, abs/1612.04426.

Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. Simple, fast, accurate intent classification and slot labeling. *arXiv preprint arXiv:1903.08268*.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the Workshop on NLP for Conversational AI*. Data available at github.com/PolyAI-LDN/conversational-datasets.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401.

Jason S. Kessler. 2017. Scattertext: a browser-based tool for visualizing how corpora differ. In *Proceedings of ACL-2017 System Demonstrations*, Vancouver, Canada. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE.

Hung Le, Truyen Tran, Thin Nguyen, and Svetha Venkatesh. 2018a. Variational memory encoder-decoder. In *Advances in Neural Information Processing Systems*, pages 1508–1518.

Hung Le, Truyen Tran, Thin Nguyen, and Svetha Venkatesh. 2018b. Variational memory encoder-decoder. In *NeurIPS*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*.

Jiwei Li, Will Monroe, Alan Ritter, Daniel Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *EMNLP*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Daniel Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *EMNLP*.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli elikyilmaz. 2017b. End-to-end task-completion neural dialogue systems. In *IJCNLP*.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the state of the art of evaluation in neural language models. *CoRR*, abs/1707.05589.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing lstm language models. *CoRR*, abs/1708.02182.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Z. Hakkani-Tür, Xiaodong He, Larry P. Heck, Gökhan Tür, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:530–539.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 234–239. IEEE.

Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In *NAACL-HLT*.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *Ninth European conference on speech communication and technology*.

Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C. Courville. 2016a. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI*.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2016c. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and William B. Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *HLT-NAACL*.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. *arXiv preprint arXiv:1605.07669*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan R. Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Xinnuo Xu, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *EMNLP*.

Kang Min Yoo, Youhyun Shin, and Sang goo Lee. 2019. Data augmentation for spoken language understanding via joint variational generation. *CoRR*, abs/1809.02305.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101:1160–1179.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive dialogue state tracker. *CoRR*, abs/1805.09655.