# Retrieval-Based Goal-Oriented Dialogue Generation

Ana Valeria González[1], Isabelle Augenstein[1], and and Anders Søgaard[1, 2]

[1]University of Copenhagen Department of Computer Science, [2]Google Research
{ana,augenstein,soegaard}@di.ku.dk ,

## Abstract

Most research on dialogue has focused either on dialogue generation for open-ended chit chat or on state tracking for goal-directed dialogue. In this work, we explore a hybrid approach to goal-oriented dialogue *generation* that combines retrieval from past history with a hierarchical, neural encoder-decoder architecture. We evaluate this approach in the customer support domain using the Multiwoz dataset (Budzianowski et al., 2018). We show that adding this retrieval step to a hierarchical, neural encoder-decoder architecture leads to significant improvements, including responses that are rated more appropriate and fluent by human evaluators. Finally, we compare our retrieval-based model to various semantically conditioned models explicitly using past dialog act information, and find that our proposed model is competitive with the current state of the art (Chen et al., 2019), while not requiring explicit labels about past machine acts.

## 1 Introduction

Dialogue systems have become a very popular research topic in recent years with the rapid improvement of personal assistants and the growing demand for online customer support. However, research has been split in two subfields (Chen et al., 2017): models presented for generation of open-ended conversations (Serban et al., 2015; Li et al., 2017a; Shibata et al., 2009; Sugiyama et al., 2013; Ritter et al., 2011) and work on solving goal-oriented dialogue through dialogue management pipelines that include dialogue state tracking and dialogue policy (Henderson et al., 2013; Ren et al., 2013; Sun et al., 2014; Zhao and Eskenazi, 2016; Mrkšić et al., 2016; Yoshino et al., 2016; Ren et al., 2018; Bingel et al., 2019).

Dialogue state tracking has often been limited to detection of user intention, as well as learning a dialogue policy to determine what actions the system should take based on the detected user intent. Dialogue generation for open ended conversation, in contrast, has largely relied on transduction architectures originally developed for machine translation (MT) (Shang et al., 2015a; Zhang et al., 2018; Wen et al., 2018). Such architectures offer flexibility because of their ability to encode an utterance into a fixed-sized vector representation, and decoding it into a variable length sequence that is linguistically very different from the input utterance. However, MT-based approaches often lack the ability to encode the context in which the current utterance occurs. This can lead to repetitive and meaningless responses (Li et al., 2015; Lowe et al., 2017; Wen et al., 2018).

This observation has led researchers to extend simple encoder-decoder models to include context in order to deal with generation of larger structured texts such as paragraphs and documents (Li et al., 2015; Serban et al., 2016, 2017). Many of these models work by encoding information at multiple levels, i.e., using both a context encoder and a last-utterance encoder, passing both encodings to a decoder that predicts the next turn. Such hierarchical methods have proven to be useful for open-ended chit chat, but were not designed for goal-oriented dialogue, where responses need not only be coherent, but also relevant.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

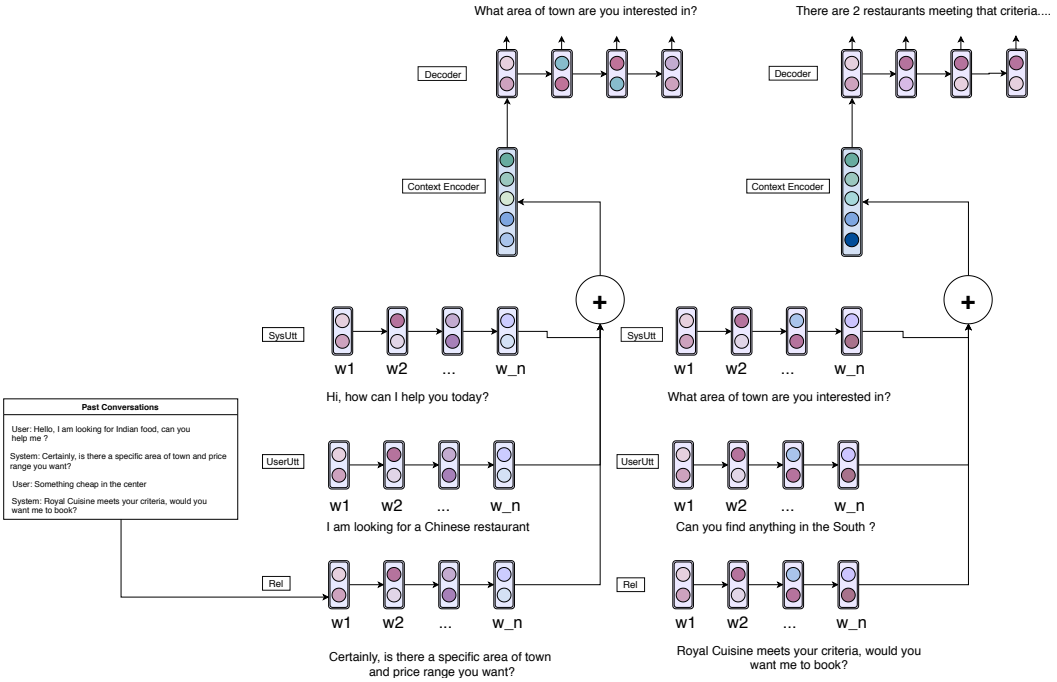


Figure 1: Our model is similar to HRED (Sordoni et al., 2015), we include an utterance encoder, a context encoder and a decoder, however, unlike HRED, our model include a simple, yet effective retrieval step used to condition the decoder to generate responses that are more appropriate for a specific domain and context.

In *goal-directed dialogue generation*, there is often one (context-dependent) right answer to a question (e.g., *How many types of insurance do you offer?*); in chit-chat, there are many good answers to questions (e.g., *What do you want to talk about today?*). We therefore hypothesize that in personal assistants and customer support, it is beneficial to increase the inductive bias of the dialogue generation model and its dependency on past conversations in order to keep responses relevant. We do so by designing a novel, hybrid dialogue generation model that conditions decoding on retrieved examplars from past history.

Although retrieval approaches to dialogue generation have been introduced before, they have typically been used to add more variety to the kind of answers the model can generate in open ended conversations (Ritter et al., 2011; Weston et al., 2018). Our model, in contrast, is designed for the purpose of goal oriented dialogue in the customer support domain. It is a hierarchical neural model with an information retrieval component that *retrieves the most informative prior turns and conditions on those*. This simple approach increases inductive bias and alleviates problems arising from long-context dependencies.

We show that an information retrieval step leads to improvements over traditional dialogue generation models intended for open ended chit chat when evaluated on BLEU and different embedding metrics. In addition, we evaluate our generated responses based on the Request/Inform success rate used in dialogue state tracking, and again, we show performance close to the state-of-the-art model which contrary to our proposed model, makes use of annotated labels. Finally, based on our human evaluations, we show that a simple retrieval step leads to system responses that are more fluent and appropriate than the responses generated by the hierarchical encoder-decoder model.

## 2 Model Description

We want to adapt commonly used models for dialogue generation to the task of goal-oriented dialogue by providing a simple way of integrating past examples in a context-aware dialogue generation model. We extend the Hierarchical Recurrent Encoder-Decoder (HRED) model presented by Sordoni et al.

(2015) for query suggestion, subsequently adopted for dialogue by Serban et al. (2016), which has shown to be a strong baseline for the task of dialogue generation. In line with previous research, we consider a dialogue $D$ between two speakers composed of $n$ utterances so that $D = [U_1, ..., U_n]$ and each utterance $U_i$ composed of $k_i$ tokens so that $U_n = [t_{i,1}, t_{i,2}, ..., t_{i,k_i}]$. Each token $t_{i,k_i}$ represents a word from a set vocabulary. Preliminary experiments showed that limiting contexts to three utterances gave the best results, therefore all results presented use $n = 3$. This is also in line with most previous work on dialogue generation and classification (Serban et al., 2016).

### 2.1 HRED

HRED (Sordoni et al., 2015) consists of an encoder RNN, which maps an utterance to a vector, a context RNN, which summarizes the dialogue history by keeping track of the previous hidden states, and a decoder RNN, which decodes the hidden state of the context RNN. Given a dialogue consisting of three utterances – a system response, a user response, and a second system response, $\langle s_1, u, s_2 \rangle$ – the goal is to predict the system utterance $s_2$ given the context $\langle s_1, u \rangle$. The utterance encoder takes in a single utterance and outputs a vector representation . The representations of each utterance are concatenated so that we end up with an array that is then fed to the context encoder. The context encoder outputs a global context that is fed into the decoder. Just as in previous work (Serban et al., 2017, 2016, 2015; Shen et al., 2017), we use GRUs (Cho et al., 2014) for our encoder, context and decoder RNNs. All modules share parameters.

### 2.2 Exemplar-HRED

In this study, we want to enhance HRED with a simple yet efficient information retrieval step. As already mentioned, similar approaches have been presented with the goal of incorporating factual information into open-ended conversations (Weston et al., 2018), to add variety and more topics to the conversation.

We hypothesize that using exemplar information is also beneficial for multi-domain goal-oriented systems. More specifically, we want to be able to inform our generation model about previous responses to similar utterances, biasing it towards past responses. For each user utterance, we extract the ten most similar past user utterances from the training set using approximate nearest neighbor search (Indyk and Motwani, 1998). We approximate a point $p \in S$ by specifying some error margin $\epsilon > 0$ so that $dist(p, q) \leq (1 + \epsilon)(dist(p*, q))$, where $p*$ is the real nearest neighbor. Because we use approximate search, we rerank the retrieved utterances using a feed-forward ranking model, introduced in Gonzalez et al. (2018). Their ranking model is a multi-task model, which relies on simple textual similarity measures combined in a multi-layered perceptron architecture. The model nevertheless achieves state-of-the-art performance on question relevancy ranking. In the end, we take the top user utterance, and return its response as the example to be used in our model.

For goal-oriented dialogue generation, our proposed model uses the same architecture as the HRED baseline, however, we include an additional RNN, which encodes the top example response. We feed the representation of the example RNN context into the context RNN and feed this representation into the decoder. Just as in the baseline model, the utterance encoder outputs a vector representation. Additionally, we encode the exemplar into a vector using the example encoder. The representations of each utterance are concatenated so that we end up with an array that includes dialogue context and exemplar information, all of which is then fed to the context encoder. The global context is then fed into the decoder.

For all experiments, we use the MultiWoz dataset for goal oriented dialogue (Budzianowski et al., 2018), which is described in the next section. Our model uses the Adam optimizer (Kingma and Ba, 2014) for all encoders. All our encoders are one layer RNNs. In addition, we use a dropout rate of 0.3, and a learning rate of 0.001. We set a maximum of 50 epochs, however, we use early stopping with a patience of 10. Most of our models converge by epoch 30. We use greedy search to generate the response during testing. More implementation details as well as our predicted utterances for each system can be found in the link provided.[1]

[1] `https://github.com/anavaleriagonzalez/exemplar_dialog`

## 3 Experiments

**Dataset** We use the MultiWoz dialogue corpus (Budzianowski et al., 2018), which consists of 10,438 dialogues spanning several domains and annotated with dialogue states and acts. We train on 8,438 dialogues, and use 1000 dialogues for development and 1000 dialogues for testing. Although the data is primarily intended for dialogue state tracking and learning a dialogue policy, Budzianowski et al. (2018) also mention its potential as a benchmark for end-to-end dialogue due to the fact that it contains about 115k turns in total, which is larger than many *structured* dialogue corpora available. This makes it a good choice for hybrid approaches in generation and goal oriented dialogue. The MultiWOZ dataset is also more difficult than the current benchmarks for goal oriented dialogue, as it spans about 7 different customer support domains and conversations are not limited to a single domain. In line with previous work in goal oriented dialog, we delexicalize the utterances to remove phone numbers, reference numbers and train ids. As opposed to other studies, we only delexicalize these three slots since these were significantly increasing the size of the vocabulary. For delexicalizing, we use the ontology provided with the data and replace the value with the slot names using regular expressions. We do not delexicalize times, prices, postcodes and distinct names of restaurants and hotels. This also makes our generation task more difficult.

**Baselines** This study is concerned with exploring simple ways of providing our dialogue generation model with enough inductive bias in order to generate fluent and on-topic responses for goal oriented dialogue. Encoder-decoder architectures work well when it comes to providing generic fluent responses for open ended conversations, however, in goal-oriented dialogue, it is also necessary for the system to remain on topic. Additionally, when training a single model on different domains, this becomes more difficult. The original HRED model (Sordoni et al., 2015; Serban et al., 2017) adapted for dialogue performs well when it comes to open ended conversations as well as when trained on very large corpora (millions of utterances) (Lowe et al., 2015; Lison and Tiedemann, 2016). In our setup however, we train the HRED model using a smaller dataset containing goal oriented dialogues in 7 different domains. We compare this model with our proposed exemplar-based model. In addition, for the BLEU metric we include the results of a transformer model(Vaswani et al., 2017) that uses dialogue context to condition the decoder as well as a LSTM that uses dialogue context and incorporates belief state and KB results as additional inputs (Budzianowski et al., 2018).

| **Metric** | **HRED** | **Exemplar-HRED** |
|---|---|---|
| BLEU | 23.6 | **24.1** |
| Vector Extrema | 59.0 | **65.0** |
| Average Embedding Similarity | 93.0 | **95.0** |
| Greedy Matching | 23.1 | **24.0** |
| Human Eval- Fluency | 0.19 | **0.58** |
| Human Eval- Appropriateness | 0.14 | **0.59** |

Table 1: The results of our dialogue generation experiments comparing HRED Sordoni et al. (2015); Serban et al. (2016) to our proposed exemplar-based model. We present resultd for standard metrics used in dialogue generation. For all the metrics we observe improvements over the strong baseline, with our best improvement of 6 percent in the vector extrema metric

## 4 Results

Overall, we found that in most cases, our model leads to significant improvements over the standard metrics (Liu et al., 2016); see Table 1 for the results. Although we are tackling goal-oriented dialogue, traditional metrics for goal oriented dialogue rely on human-generated supervision i.e. slot-value pair labels or dialogue act labels. Word overlap metrics such as the ones used for machine translation are often used to evaluate the quality of dialogue generation, however, these standard metrics tend to have very weak correlation with human judgment. In any case, we include these, as well as word embedding metrics for comparison. For the standard metrics, we use the evaluation scripts from

(Serban et al., 2016) [2]. We observe that the retrieval model is consistently better across all scenarios and metrics. In addition to these metrics, we assess our performance using the dialogue success metrics typically used in belief tracking (Budzianowski et al., 2018). We briefly explain these metrics further.

**BLEU** BLEU (Papineni et al., 2002) is typically used for machine translation and has subsequently been used to evaluate the performance of many dialogue generation systems (Galley et al., 2015; Serban et al., 2017, 2016). BLEU analyzes co-occurrences of n-grams in a reference sequence and a hypothesis. For all datasets, we see improvements with BLEU. It uses a modified precision to account for the differences in length between reference and generated output. Given a reference sentence $s$ and a hypothesis sentence $\hat{s}$, we can denote the n-gram precision $P_n(s, \hat{s})$ as: $P_n(s, \hat{s}) = \frac{\sum_q min(h(q,s),(h(q,\hat{s}))}{\sum_q h(q,s)}$ where q is the index of all possible n-grams, and h(q,s) is the number of n-grams in s.

**Average Word Embedding Similarity** We follow Liu et al. (2016) and obtain the average embedding $e_s$ for the reference sentence $s$ by averaging the word embeddings $e_w$ for each token $w$ in $s$. We do the same for the predicted output $\hat{s}$ and obtain the final similarity score by computing cosine similarity of the two resulting vectors. Again, Exemplar-HRED is consistently superior yielding almost a 2 percent improvement over the best baseline model.

**Vector Extrema** We also compute the cosine similarity between the vector extrema of the reference and the hypothesis, again following Liu et al. (2016). The goal of this metric as described by the authors is to consider informative words rather than common words, since the vectors for common words will tend to be pulled towards the zero vector. Our exemplar model achieves the largest improvement for this metric, with a gain of 6 percent over the baseline model.

**Greedy Matching** In greedy matching (Liu et al., 2016), given two sequences $s$ and $\hat{s}$, each token $w \in s$ is matched with each token $\hat{w} \in \hat{s}$ by computing the cosine similarity of the corresponding word embeddings $emb_w$ and $emb_{\hat{w}}$. The local match $g(s, \hat{s})$ is the word embedding with the maximum cosine similarity. We compute in both directions and the total score is: $G(s, \hat{s}) = \frac{g(s,\hat{s})+g(\hat{s},s)}{2}$ This metric is used to favour key words. Our best model shows only small improvements on this metric.

**Human Evaluation** In addition to the previously mentioned standard metrics, we also evaluate the performance of our baseline and the exemplar-based models using human evaluations. We extract 100 baseline and exemplar model system responses at random. We ask the 7 evaluators to 1) pick the response that is more fluent and grammatically correct and 2) pick the response that achieves the goal given the context of the conversation. We provide the context of System and User utterances, and ask the evaluators to pick one of 4 options: 1) the output of the baseline, 2) the output of the exemplar model, 3) both, 4) none. The order of the options was shuffled.

Overall, we found that when it came to fluency, the evaluators perceived that 58% of the time, the exemplar response was better. The baseline beat the exemplar based response for 19 percent of the evaluated dialogs and the rest of the dialogs either both or none were picked. For appropriateness we see a similar pattern. Evaluators perceived the response produced by the exemplar model as the more appropriate one given the context, for 59 percent of the evaluated dialogs. The baseline beat the proposed model only 14 percent of the time. These results can also be found on Table 1.

**Dialogue success: inform/request** Traditional goal-oriented dialogue systems based on prediction of slots and dialogue acts are typically evaluated on the accuracy of predicting these as labels, as well as their success at the end of the dialogue. Dialogue success is measured by how many correct inform/request slots a model can generate in a conversation in comparison to the ground truth. An inform slot is one that provides the user with a specific item for example the inform slots "food" and "area" i.e. (food="Chinese", area="center") informs the user that there is a Chinese restaurant in the center. On the other hand, a request slot is a slot that specifies what information is needed for the system to achieve the user goal. For example, for booking a train, the system needs to know the departure location and the destination. The slots "departure" and "destination" would be the request slots in this case.

[2] https://github.com/julianser/hed-dlg-truncated/tree/master/Evaluation

| | Model | Inform | Request | BLEU |
|---|---|---|---|---|
| No act | 3-layer Transformer (Vaswani et al., 2017) | 71.1 | 59.9 | 19.1 |
| | HRED | 60.4 | 44.5 | 23.6 |
| | Exemplar-HRED | 77.6 | 70.1 | 24.1 |
| Act | LSTM (Budzianowski et al., 2018) | 71.2 | 60.2 | 18.8 |
| | SC-LSTM (Wen et al., 2015) | 74.5 | 62.5 | 20.5 |
| | RL(Mehri et al., 2019) | 82.7 | 72.1 | 16.3 |
| | HDSA-predicted | 82.9 | 68.9 | 23.6 |
| | HDSA-groundtruth | 87.9 | 78.0 | 30.4 |

Table 2: Inform/request results divided into two section. The top models do not make use of any past dialog acts to condition the decoder to generate a response. The models at the bottom use dialog acts, and belief state in order to generate better responses

For goal-oriented generation, many of the models evaluated using the Inform/Request metrics have made use of structured data to semantically condition the generation model in order to generate better responses (Wen et al., 2015; Chen et al., 2019; Budzianowski et al., 2018; Zhao et al., 2019; Mehri et al., 2019) . Wen et al. (2015) proposed to encode each individual dialog act as a unique vector and use it as an extra input feature, in order to influence the generated response. This method has been shown to work well when tested on single domains where the label space is limited to a few dialog acts. However, as the label space grows, using a one-hot encoding representation of the dialog act is not scalable. To deal with this problem, Chen et al. (2019) introduced a semantically conditioned generation model using Hierarchical Disentangled Self-Attention (HDSA). This model deals with the large label space by representing dialog acts using a multi-layer hierarchical graph that merges cross-branch nodes. For example, the distinct trees for HOTEL-RECOMMEND-AREA and ATTRACTION-RECOMMEND-AREA can be merged at the second and third levels sharing semantic information about actions and slots but maintaining specific information about the domains separate. This model achieves the state-of-the-art result for generation in both BLEU and Inform/Request metrics.

In order to assess how our simple approach compares to models explicitly using dialog acts as inputs, we compute the inform/request accuracy and compare to the state-of-the-art (Chen et al., 2019) as well as other baseline models. In addition, we include comparisons to Mehri et al. (2019) which uses reinforcement learning as well as past machine act information and achieve a strong performance.

Chen et al. (2019) present experiments conditioning both on predicted acts as well as ground truth past acts. We include both of these as well as the performance of our baseline and proposed model in Table 2. We divide the results into models using act information to condition the language generation and models that do not.

## 5 Discussion

As shown in Table 1, our simplest proposed model achieved the largest improvements over the baseline when it came to the average embedding similarity and vector extrema similarity. As it is hard to interpret what the difference in performance of each model is based on standard dialogue metrics we examined the output to spot the major differences in response generation of our proposed models versus the baseline. We looked at the responses generated by our proposed models that had the highest score of these metrics and compared to the response generated by the baseline for that same dialogue. Overall we found that the baseline models tend to generate responses containing slots and values for the wrong domain. In addition, by examining the outputs we could see that the vector extrema metric is very sensitive when it comes to slight differences in the references and prediction. We found that this metric was more indicative of model performance than embedding similarity. We present some example outputs in Table 3.

As mentioned earlier, from manual inspection of the outputs we observed that the the exemplar model is able to stay within the correct domain of the conversation and returns information within that domain that is more appropriate given the conversation context. This was confirmed by the

| Gold Response | Exemplar-HRED | HRED |
|---|---|---|
| your booking was successful and your reference number is hotel-reference . is there anything else i can help you with ? | your booking was successful . your reference number is hotel-reference . is there anything else i can help with ? | the booking was successful . the table will be reserved for 15 minutes . reference number is : restaurant-reference . |
| **Avg emb score** | 0.99 | 0.93 |
| **Vector Extrema** | 0.99 | 0.65 |
| booking was successful . the table will be reserved for 15 minutes . your reference number is : restaurant-reference . | your reservation was successful . the table will be reserved for 15 minutes . reference number is : restaurant-reference . | the train-trainid is train-trainid and arrives at stansted airport at the airport at 15:08 . would you like a booking |
| **Avg emb score** | 0.99 | 0.91 |
| **Vector Extrema** | 0.90 | 0.56 |
| there is no option for me to book it . is there anything else i can help you with ? | certainly ! the reference number is restaurant-reference . can i help you with anything else ? | the train-trainid leaves at 09:01 and arrives in cambridge by 09:52 . would that work ? |
| **Avg emb score** | 0.99 | 0.92 |
| **Vector Extrema** | 0.74 | 0.51 |
| yes , the address is saint andrew 's street and the phone number is hotel-phone . is there anything else i can assist you with today ? | the address is saint andrew 's street , and the phone number is hotel-phone . can i help you with | your taxi is booked for you . the contact number is taxi-phone . is there anything else i can help |
| **Avg emb score** | 0.99 | 0.98 |
| **Vector extrema** | 0.98 | 0.75 |
| the travel time is 88 minutes . | the travel time is 79 minutes . | what time would you like to travel ? |
| **Avg emb score** | 0.99 | 0.88 |
| **Vector Extrema** | 0.97 | 0.59 |
| where will you leave from ? | sure , when would you like to leave ? | have a nice stay . |
| **Avg emb score** | 0.98 | 0.95 |
| **Vector Extrema** | 0.94 | 0.74 |

Table 3: Examples of responses generated by both the baseline and our proposed model. By examining the outputs, it becomes noticeable that the baseline model tends to generate responses that are not precise about the current domain of the conversation (hotel, taxi booking, trains, restaurant, etc). The Exemplar-HRED model on the other hand becomes more accurate in responding for the correct domain. Here we present a few examples of responses given the same context, and provide their scores (similarity to reference) for comparison

human evaluations and also the Inform/Request metrics. When comparing the performance of the exemplar-based model to models that do not use information about past acts to condition the decoder, we observe that including a simple retrieval step leads to very large gains in the success of providing the inform/request slots.. In addition, the exemplar model performs better than Budzianowski et al. (2018), which uses information of the belief state of the conversation as extra features. More interestingly, our proposed model performs better than the state-of-the-art when it comes to providing the request slots. It also outperforms this same model when evaluated on BLEU; however, it still falls behind the state-of-the-art when it comes to providing inform slots. Overall, we find that our model remains competitive without requiring turn labels.

## 6 Related Work

Dialogue generation has relied on transduction architectures originally developed for machine translation (MT) (Shang et al., 2015a; Zhang et al., 2018; Wen et al., 2018). Open domain dialogue systems aim to generate fluent and meaningful responses, however this has proven a challenging

task. Most systems are able to generate coherent responses that are somewhat meaningless and at best entertaining (Lowe et al., 2017; Wen et al., 2018; Serban et al., 2016). Much of the research on dialogue generation has tried to tackle this problem by predicting an utterance based on some dialogue history (Vinyals and Le, 2015; Shang et al., 2015b; Luan et al., 2016; Serban et al., 2016). We extend such an architecture to also include past history, in order to avoid generating too generic responses.

Most research on goal-oriented dialogue has focused almost exclusively on dialogue state tracking and dialogue policy learning (Sun et al., 2014, 2016; Li et al., 2017b; Henderson, 2015; Henderson et al., 2014; Rastogi et al., 2017; Mrkšić et al., 2016; Yoshino et al., 2016; Bingel et al., 2019). Dialogue state tracking consists of detecting the user intent and tends to rely on turn-level supervision and a preset number of possible slot and value pairs which limits the flexibility of such chatbots, including their ability to respond to informal chit chat, as well as transferring knowledge across domains. There has been some work in the past few years that has attempted to address these problems by introducing methods that focus on domain adaptation as well as introducing new data to make this task more accessible (Rastogi et al., 2017, 2019; Budzianowski et al., 2018; Mrkšić et al., 2015; Bingel et al., 2019). Recent approaches have also introduced methods for representing slot and value pairs that do not rely on a preset ontology (Ren et al., 2018; Mrkšić et al., 2017), in an attempt to add flexibility. In our research, we acknowledge the importance of this added flexibility in goal oriented dialogue and propose a method for generating goal oriented responses without having turn level supervision.

The idea of combining text generation with past experience has been explored before. White and Caldwell (1998) used a set of hand crafted examples in order to generate responses through templates. More recently, Song et al. (2016) also explored a hybrid system with an information retrieval component, but their system is very different: It uses a complex ranking system at a high computational cost, requires a post-reranking component to exploit previous dialogue turns (about half of which are copied over as predictions), and they only evaluate their system in a chit-chat set-up, reporting only BLEU scores. In a similar paper, (Weston et al., 2018) tried to move away from short generic answers in order to make a chit-chat generation model more entertaining by using an information retrieval component, to introduce relevant facts. In addition, a similar method was recently shown to improve other generation tasks such as summarization. In Subramanian et al. (2019), the authors show that a simple extractive step introduces enough inductive bias for an abstractive summarization system to provide fluent yet precise summaries. In contrast to these works, we integrate a retrieval based method with a context-aware neural dialogue generation model in order to introduce relevant responses in a goal oriented conversation.

## 7 Conclusion

In this study, we have experimented with a simple yet effective way of conditioning the decoder in a dialogue generation model intended for goal oriented dialogue. Generating fluent *and* precise responses is crucial for creating goal-oriented dialogue systems, however, this can be a very difficult task; particularly, when the system responses are dependent on domain-specific information. We propose adding a simple retrieval step, where we obtain the past conversations that are most relevant to the current one and condition our decoder on these. We find that this method not only improves over multiple strong baseline models on word overlap metrics, it also performs better than the state-of-the-art on BLEU and achieves competitive performance for inform/request metrics without requiring dialog act annotations. Finally, by inspecting the output of the baseline versus our proposed model and through human evaluations, we find that a great advantage of our model is its ability to produce responses that are more fluent and remain on topic.

## 8 Acknowledgements

We thank the reviewers for their valuable comments. Ana Valeria González and Anders Søgaard are funded by a Google Focused Research Award, a Facebook Research Award, as well as by Innovation Fund Denmark.

## References


Joachim Bingel, Victor Petrén Bach Hansen, Ana Valeria Gonzalez, Paweł Budzianowski, Isabelle Augenstein, and Anders Søgaard. 2019. Domain Transfer in Dialogue Systems without Turn-Level Supervision. In *3rd Conversational AI Workshop at NeurIPS 2019*.

Paweł 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. *arXiv preprint arXiv:1905.12866*.

Kyunghyun Cho, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*.

Ana Gonzalez, Isabelle Augenstein, and Anders Søgaard. 2018. A strong baseline for question relevancy ranking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4810–4815.

Matthew Henderson. 2015. Machine learning for dialog state tracking: A review. In *Proc. of The First International Workshop on Machine Learning in Spoken Language Processing*.

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

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.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA. ACM.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR*.

Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1106–1115.

Jiwei Li, Will Monroe, Tianlin Shi, Sèbastien Jean, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

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

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas A.-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *ACL*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Yi Luan, Yangfeng Ji, and Mari Ostendorf. 2016. Lstm based conversation models. *arXiv preprint arXiv:1603.09457*.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog.

N Mrkšić, DO Séaghdha, B Thomson, M Gašić, PH Su, D Vandyke, TH Wen, and S Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. In *ACL-IJCNLP 2015-53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 2, pages 794–799.

Nikola Mrkšić, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2016. Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1777–1788.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Abhinav Rastogi, Dilek Hakkani-Tür, and Larry Heck. 2017. Scalable multi-domain dialogue state tracking. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 561–568. IEEE.

Pushpendre Rastogi, Arpit Gupta, Tongfei Chen, and Lambert Mathias. 2019. Scaling multi-domain dialogue state tracking via query reformulation. *arXiv preprint arXiv:1903.05164*.

Hang Ren, Weiqun Xu, Yan Zhang, and Yonghong Yan. 2013. Dialog state tracking using conditional random fields. In *Proceedings of the SIGDIAL 2013 Conference*, pages 457–461.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR, abs/1507.04808*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015a. Neural responding machine for short-text conversation. In *ACL*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015b. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1577–1586.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 504–509.

Masahiro Shibata, Tomomi Nishiguchi, and Yoichi Tomiura. 2009. Dialog system for open-ended conversation using web documents. *Informatica*, 33(3).

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562. ACM.

Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher Pal. 2019. On extractive and abstractive neural document summarization with transformer language models. *arXiv preprint arXiv:1909.03186*.

Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2013. Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. In *Proceedings of the SIGDIAL 2013 Conference*, pages 334–338.

Kai Sun, Lu Chen, Su Zhu, and Kai Yu. 2014. A generalized rule based tracker for dialogue state tracking. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 330–335. IEEE.

Kai Sun, Su Zhu, Lu Chen, Siqiu Yao, Xueyang Wu, and Kai Yu. 2016. Hybrid dialogue state tracking for real world human-to-human dialogues. In *INTERSPEECH*, pages 2060–2064.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals and Quoc V Le. 2015. A neural conversational model.

Haoyang Wen, Yijia Liu, Wanxiang Che, Libo Qin, and Ting Liu. 2018. Sequence-to-sequence learning for task-oriented dialogue with dialogue state representation. In *COLING*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92.

Michael White and Ted Caldwell. 1998. Exemplars: A practical, extensible framework for dynamic text generation. *Natural Language Generation*.

Koichiro Yoshino, Takuya Hiraoka, Graham Neubig, and Satoshi Nakamura. 2016. Dialogue state tracking using long short term memory neural networks. In *Proceedings of Seventh International Workshop on Spoken Dialog Systems*, pages 1–8.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. pages 1–10.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. *arXiv preprint arXiv:1902.08858*.