DEEPCOPY: Grounded Response Generation with Hierarchical Pointer Networks

Semih Yavuz*
University of California, Santa Barbara
syavuz@cs.ucsb.edu

Abhinav Rastogi
Google AI
abhirast@google.com

Guan-lin Chao*
Carnegie Mellon University
guanlinchao@cmu.edu

Dilek Hakkani-Tür*
Amazon Alexa AI
dilek@ieee.org

Abstract

Recent advances in neural sequence-to-sequence models have led to promising results for several language generation-based tasks, including dialogue response generation, summarization, and machine translation. However, these models are known to have several problems, especially in the context of chit-chat based dialogue systems: they tend to generate short and dull responses that are often too generic. Furthermore, these models do not ground conversational responses on knowledge and facts, resulting in turns that are not accurate, informative and engaging for the users. These indeed are the essential features that dialogue response generation models should be equipped with to serve in more realistic and useful conversational applications. Recently, several dialogue datasets accompanied with relevant external knowledge [29, 5] have been released to facilitate research into remedying such issues encountered by traditional models by resorting to this additional information. In this paper, we propose and experiment with a series of response generation models that aim to serve in the general scenario where in addition to the dialogue context, relevant unstructured external knowledge in the form of text is also assumed to be available for models to harness. Our approach extends pointer-generator networks [18] by allowing the decoder to hierarchically attend and copy from external knowledge in addition to the dialogue context. We empirically show the effectiveness of the proposed model compared to several baselines including [6, 29] on CONVAI2 challenge.

1 Introduction

Recently, deep neural networks have achieved state-of-the-art results in various tasks including computer vision, natural language processing, and speech processing. Specifically, neural sequence-to-sequence models [22, 2] have led to great progress in important downstream NLP tasks like text summarization [17, 4, 13, 12, 18, 23, 30], machine translation [11, 2, 3, 22], and reading comprehension [28]. However, achieving satisfactory performance on dialogue tasks still remains an open problem. This is because dialogues can have multiple valid responses with varying semantic content. This is vastly different from the aforementioned tasks, where the generation is more conveniently and uniquely constrained by the input source.

Although neural models appear to generate meaningful responses when trained with sufficiently large datasets in the chit-chat setting, such generic chit-chat models reveal several weaknesses that are reported by previous research [19, 26] as well. Most common problems include inconsistency

*Work done at Google AI.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

in personality, dull and generic responses, and unawareness of long-term dialogue context. The inconsistency in personality arises because the model learns to generate the dialogues from a set of many speakers with different backgrounds, interests and even language styles. Furthermore, response generation models are usually trained using only the recent dialogue context, and therefore the model is unaware of long-term history and tend to generate outputs that lack connection with earlier turns. Such chit-chat models learn to generate generic responses such as "Really?", "Wow", "I don't know" which appear frequently in the training set but are also uninteresting.

To alleviate these limitations, we train our dialogue generation model based off CONVAI2 challenge, in a setting where the model is provided a set of relevant textual facts (speaker persona descriptions) and is allowed to harness this knowledge when generating responses in a multi-turn dialogue. To handle the personalty inconsistency issue, we ground our dialogue generation model on external knowledge facts which are a list of persona descriptions in our application [8, 29]. We explicitly use the dialogue history as memory for the model to condition on which potentially encourages a more natural conversation flow. To prevent from generating generic but dull responses, we use a hierarchical pointer network in our model such that it can copy content from two sources: current dialogue history and persona descriptions, to encourage specific and appropriate responses.

In this work, we propose a novel architecture DEEPCOPY that extends the attentional sequence-to-sequence model with a hierarchical pointer network that enables the decoder to jointly attend and copy tokens from any of the facts available as external knowledge in addition to the dialogue context (encoder input). This is achieved entirely in an end-to-end fashion through factoring the whole copy mechanism into following three hierarchies/components: (i) a token-level attention mechanism over the dialogue context to determine the probability of copying a token from the dialogue context, (ii) A hierarchical pointer network to determine the probability of copying a token from each fact, and (iii) An inter-source meta attention over the input sources *dialogue context* and *external knowledge*, which combines the two copying probabilities. Using these components, a single copying probability distribution over the unique tokens appearing in the model input is computed exploiting the well-defined hierarchy among them. In addition, the model is equipped with a soft switch mechanism between *copying* and *generation* modes similar to [18], which allows us to softly combine the *copying probabilities* with the decoder's *generation probabilities* over a fixed vocabulary into a final output probability distribution over an extended vocabulary. We empirically show the effectiveness of the proposed DEEPCOPY model compared to several baselines including [6, 29] on CONVAI2 challenge.

2 Related Work

Earlier work on data-driven, end-to-end approaches to conversational response generation treated the task as statistical machine translation, where the goal is to generate a response given the previous dialogue turn [16, 26]. While these studies resulted in a paradigm change compared to earlier work based on rule- and retrieval-based approaches, responses generated by these sequence-to-sequence models are not always contextually appropriate, as they do not include mechanisms to represent longer term conversation context. To tackle this problem and have a better representation of conversation context as input to generation, [19] proposed hierarchical recurrent encoder-decoder (HRED) networks. HRED combines two RNNs, one at the token level, modeling individual turns, and one at the dialogue level, inputting turn representations from the token-level RNNs. However, utterances generated by such neural response generation systems are noted to be often generic and contentless [26]. To improve the diversity and content of generated responses, HRED was later extended with a latent variable that aims to model the higher level aspects (such as topic) of the generated responses, resulting in the VHRED approach [20].

Another issue that remains with these previous approaches is that the generated dialogues mainly rely on conversation context and do not integrate knowledge. [10] extracted facts relevant to the current dialogue from a knowledge base using string matching, entity linking or named entity recognition, found an additional K entities from the knowledge base that are most relevant to the facts by a neural similarity scorer, and combined the information as context feature to the dialogue generation RNN. [6] used end-to-end memory networks to base the generated responses on knowledge, where an attention over the knowledge relevant to the conversation context is estimated, and multiple knowledge representations are included as input during the decoding of responses. In this work, we use end-to-end memory networks as one of the baselines.

Although much research has focused on response generation in a chit-chat setting, models trained on large datasets of human-human interactions of diverse speaker characteristics often tend to generate

responses which are too vague and generic (common for most speakers) or inconsistent in personality (switching between different speakers' characteristics). Therefore, persona and speaker modeling has remained a challenge for conversational agents due to the lack of publicly available dialogue datasets to this end. Recently, [29] presented the CONVAI2 challenge containing persona descriptions and over 10K real human chit-chats where speakers were required to converse based on their assigned persona. [8] learned speaker persona embeddings from a single-speaker setting (e.g. Twitter posts) or a speaker-address style (human-human conversations) to generate personalized responses given a single utterance input. Another related work [15] applies hierarchical memory network for task oriented dialog problem. In this work, we compared our model with [29] which used a memory-augmented sequence-to-sequence response generator grounded on the dialogue history and persona.

3 Model

In this section, we first set up the problem, and then briefly revisit the baseline models using memory networks [21] and pointer-generator networks [18]. Subsequently, we introduce the proposed DEEPCOPY model with hierarchical pointer network and our training process.

3.1 Problem Setup

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote the tokens in the dialogue history. The dialogue is accompanied by a set of K relevant supporting facts, where $\mathbf{f}^{(i)} = (f_1^{(i)}, f_2^{(i)}, \ldots, f_{n_i}^{(i)})$ is the list of tokens in the i-th fact. Our goal is to generate the response as a sequence of tokens $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ using the dialogue history and supporting facts. Note here that we are not interested in retrieval/ranking based models [27] which rely on a set of candidate responses. Generative models are essential for this problem because we want to incorporate content from new facts during inference which may not be present in the training set. Hence, using a predefined set of candidates may not ensure high coverage.

3.2 Baseline Models

In this section, we describe several baseline response generation models including the ones from existing work [6, 29] and the in-house ones we propose as additional baselines. We describe the individual baseline models as follows.

3.2.1 Seq2Seq

In a sequence-to-sequence model with attention [2], a sequence of input tokens is encoded using an LSTM encoder. At decoder step t, the decoder state h_t, a context vector c_t and the previous decoder output y_{t-1} are used together to output a distribution over a fixed vocabulary of tokens obtained from the training set using a non-linear function. The context vector c_t is an attention-weighted combination of the encoder outputs. In the following baseline models, we use different features as inputs to the encoder. The underlying model remains the same.

- SEQ2SEQ + NOFACT: Only the dialogue context tokens $\mathbf{x}$ are used as input to the encoder.
- SEQ2SEQ + BESTFACTCONTEXT: We select the fact $\mathbf{f}^{(c)}$ whose tokens have highest unigram *tf-idf* similarity to the dialogue context tokens. $[\mathbf{x}||\mathbf{f}^{(c)}]$ is then used as input to the encoder, where $||$ denotes concatenation.
- SEQ2SEQ + BESTFACTRESPONSE: We select the fact $\mathbf{f}^{(r)}$ whose tokens have highest unigram *tf-idf* similarity to the ground truth response. $[\mathbf{x}||\mathbf{f}^{(r)}]$ is used as input to the encoder. The aim of this experiment is to have a better understanding of the effect of fact selection on response generation, since using the ground truth for fact selection is not fair.

3.2.2 Memory Network

Our variations of Seq2Seq models described in Section 3.2.1 incorporate facts by concatenating them to the dialogue context. Memory networks [6, 29] are a more principled approach to incorporate external facts. Similar to [6], we use a context encoder to embed the context tokens $\mathbf{x}$ and obtain a list of outputs and final hidden state $u \in \mathbb{R}^d$. As outlined in [6], a fact $\mathbf{f}^{(i)}$ is embedded into key and value vectors k_i and m_i, respectively. A summary $o \in \mathbb{R}^d$ of facts is then computed as an attention weighted combination of $(m_1, m_2, \ldots, m_K)$ by conditioning on u and $(k_1, k_2, \ldots, k_K)$. We then combine the two summaries into $\hat{u} = u + o$, and use it to initialize the decoder state. We report results on the following variants:

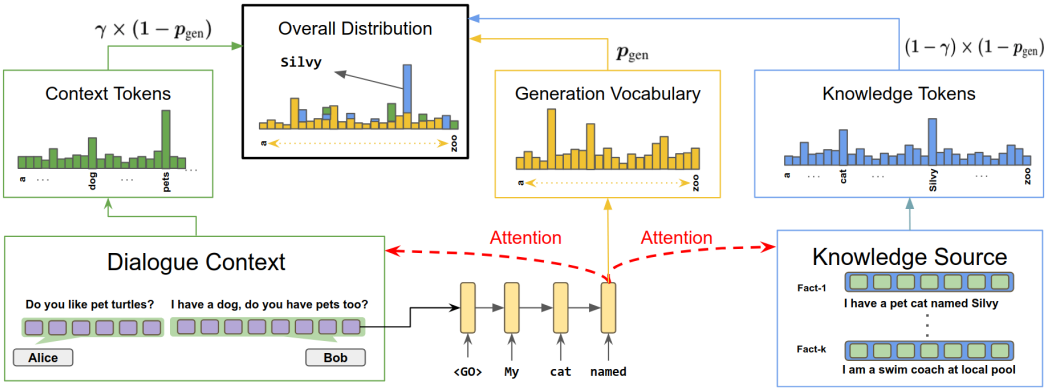

Figure 1: Overview of our proposed approach as described in Section 3.3. The decoder state d_t is used to attend over dialogue context and knowledge source to generate distributions for copying tokens from these sources. The decoder also outputs a distribution over a fixed vocabulary. The three distributions are combined to yield the final distribution over tokens at each decoder step t.

- MEMNET: This is equivalent to the model used in [6], as described above. This is essentially equivalent to a sequence to sequence model without attention at every decoder step, except using the combined summary $\hat{u}$ to initialize the decoder.
- MEMNET+CONTEXTATTENTION: At each decoder step, the decoder state attends over the encoder outputs and obtains a context vector $c_t^{(c)}$. This is equivalent to SEQ2SEQ + NOFACT model from Section 3.2.1, except using the fact summary $\hat{u}$ to initialize the decoder state.
- MEMNET+FACTATTENTION: At each decoder step, we use the decoder state to attend over the value embeddings $(m_1, m_2, \ldots, m_K)$ corresponding to facts, and obtain a context vector $c_t^{(f)}$. This model is similar to *generative profile memory network* [29], where we apply attention only on facts, and we set the decoder's initial state to the combined summary $\hat{u}$ from context and facts.
- MEMNET+FULLATTENTION: This model employs attention over both facts and dialogue context at each decoder step. The two attention modules are combined by concatenating $c_t^{(c)}$ and $c_t^{(f)}$ [31].

3.2.3 Seq2Seq with Copy Mechanism

Seq2seq models can only generate tokens present in a fixed vocabulary obtained from the training set. This can be a big drawback if we want the model to be able to generate responses containing concepts not seen in the training set. Pointer-generator network [18] extends the attentional sequence-to-sequence model [2] by employing a pointer network [25]. It has two decoding modes, copying and generating, which are combined via a soft switch mechanism, allowing it to copy tokens from source in addition to generating from vocabulary. We report the results for the following additional baselines obtained by equipping the corresponding Seq2Seq model described in Section 3.2.1 with copy mechanism: SEQ2SEQ + NOFACT + COPY, SEQ2SEQ + BESTFACTCONTEXT + COPY, SEQ2SEQ + BESTFACTRESPONSE + COPY.

3.3 DeepCopy with Hierarchical Pointer Networks

Pointer-generator network [18] can only copy tokens from the encoder input. In this section, we present our proposed DEEPCOPY model that extends pointer-generator network [18] using a novel hierarchical pointer network. Our model allows copying tokens from multiple input sources (facts $\mathbf{f}^{(i)}, 1 \leq i \leq K$), besides the encoder input (dialogue context x).

A high-level overview of the proposed approach is illustrated in Figure 1. At decoder step t, the decoder state h_t is used to attend over the dialogue context tokens and fact tokens to give a distribution over the tokens present in context and facts respectively. These distributions are then combined with the distribution output by the decoder over the fixed vocabulary to obtain the overall distribution.

Encoding a sequence. Let $\mathbf{w} = (w_1, w_2, \ldots w_n)$ be a sequence of tokens. We first obtain a trainable embedded representation of each token in the sequence and then use a LSTM cell to encode

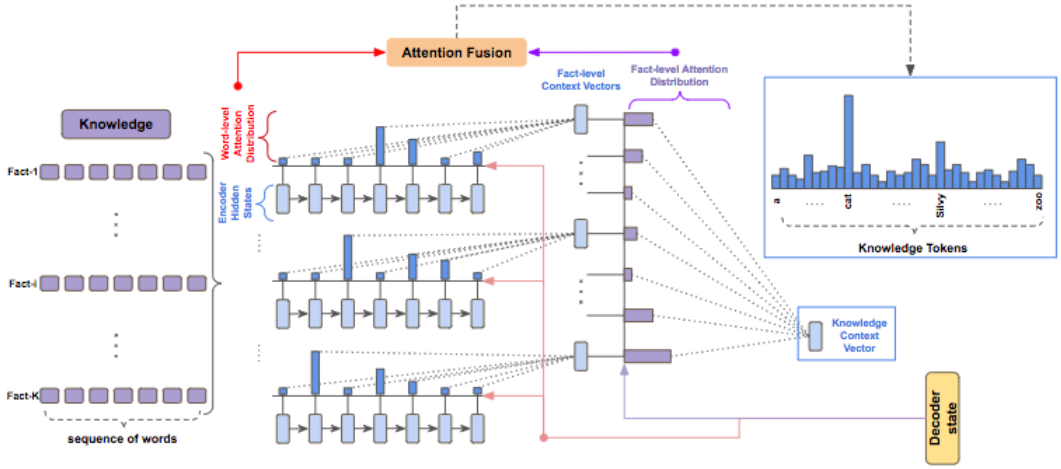

Figure 2: Illustration of hierarchical pointer network. The decoder state d_t is used to attend over tokens for each fact and also over the fact-level context vectors obtained by weighted average of token-level representations (w.r.t token-level attention weights) for each fact. The token-level attention weights are then combined with the attention distribution over facts (Equation 11) to generate the probability of copying each token in all the facts.

the sequence of embedding vectors. We define $e, \mathbf{s} = \texttt{Encode}(\mathbf{w})$, where e denotes the final state of the LSTM and $\mathbf{s} = (s_1, s_2, \ldots s_n)$ denotes the outputs of the LSTM cell at all steps.

Attention. Let $\mathbf{u} = (u_1, u_2, \ldots u_n)$ be a sequence of vectors where $u_i \in \mathbb{R}^p, 1 \leq i \leq n$ and $v \in \mathbb{R}^q$ be a conditioning vector. The attention module generates a linear combination c of elements in $\mathbf{u}$ by conditioning them on v as defined by the equations below. We define $\alpha, c = \texttt{Attention}(\mathbf{u}, v)$, where $\alpha_i \in \mathbb{R}^n$ is the weight assigned to u_i, and $c \in \mathbb{R}^p$ is a vector representation of the sequence $\mathbf{u}$ conditioned on v. In the equations below, w_1 and W_2 are parameters of appropriate dimension. In our setup, we use $p = q, w_1 \in \mathbb{R}^p$, and $W_2 \in \mathbb{R}^{p \times 2p}$.

$$e_i = w_1^T \tanh(W_2[u_i; v]) \tag{1}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)} \tag{2}$$

$$c = \sum_{i=1}^{n} \alpha_i u_i \tag{3}$$

Copying from Dialogue Context. Similar to our baseline models, we encode the dialogue context tokens $\mathbf{x}$ (Equation 4) and apply attention to the encoder outputs at a decoder step t (Equation 5). This outputs attention weights $\alpha_t^{(x)}$ and a representation of the entire context $c_t^{(x)}$. The attention weights are aggregated to obtain the distribution over context tokens $p_t^{(x)}(w)$ (Equation 6),

$$e^{(x)}, \mathbf{s}^{(x)} = \texttt{Encode}(\mathbf{x}) \tag{4}$$

$$\alpha_t^{(x)}, c_t^{(x)} = \texttt{Attention}(\mathbf{s}^{(x)}, h_t) \tag{5}$$

$$p_t^{(x)}(w) = \sum_{\{i: x_i = w\}} \alpha_{t,i}^{(x)} \tag{6}$$

Copying from Facts: Hierarchical Pointer Network. We introduce the hierarchical pointer network (Figure 2) as a general methodology for enabling token-level copy mechanism from multiple input sequences or facts. Each fact $\mathbf{f}^{(i)}$ is encoded (Equation 7) to obtain token level representations $\mathbf{s}^{(f)(i)}$ and overall representation $e^{(f)(i)}$. The decoder state h_t is used to attend over token level representations of each fact (Equation 8) and the overall representations of all the facts (Equation 9).

$$e^{(f)(i)}, \mathbf{s}^{(f)(i)} = \texttt{Encode}(\mathbf{f}^{(i)}) \tag{7}$$

$$\alpha_t^{(f)(i)}, c_t^{(f)(i)} = \texttt{Attention}(\mathbf{s}^{(f)(i)}, h_t) \tag{8}$$

$$\beta_t, c_t^{(f)} = \texttt{Attention}([c_t^{(f)(1)}, \ldots, c_t^{(f)(K)}], h_t) \tag{9}$$

$$p_t^{(f)}(w) = \sum_{j=1}^{K} p_t^{(f)}(\mathbf{f}^{(j)}) \cdot p_t^{(f)}(w|\mathbf{f}^{(j)}) = \sum_{j=1}^{K} \beta_{t,j} \sum_{\{l:f_l^{(j)}=w\}} \alpha_{t,l}^{(f)(j)} \tag{10}$$

Inter-Source Attention Fusion We now present the mechanism to fuse the two distributions $p_t^{(x)}(w)$ and $p_t^{(f)}(w)$ representing the probabilities of copying tokens from dialogue context and facts respectively. We use the decoder state h_t to attend over dialogue context representation $c_t^{(x)}$ and overall fact representation $c_t^{(f)}$ (Equation 11). The resulting attention weight $\gamma_t' = [\gamma_t, 1-\gamma_t]$ is used to combine the two copying distributions as shown in Equation 12.

$$\gamma_t, c_t = \texttt{Attention}([c_t^{(x)}, c_t^{(f)}], h_t) \tag{11}$$

$$p_t^{\text{copy}}(w) = \gamma_t \, p_t^{(x)}(w) + (1-\gamma_t) \, p_t^{(f)}(w) \tag{12}$$

Similar to Seq2Seq models, the decoder also outputs a distribution p_t^{vocab} over the fixed training vocabulary at each decoder step using the overall context vector c_t and decoder state h_t. Having defined the copy probabilities p_t^{copy} for tokens that appear in the model input, either the dialogue context or the facts in external knowledge source, we combine p_t^{vocab} and p_t^{copy} using the mechanism outlined in [18], except we use c_t defined in Equation 11 as the context vector instead.

To better isolate the effect of copying, a key component of the proposed DEEPCOPY model, we also conduct experiments with MULTISEQ2SEQ model that incorporates the knowledge facts in the same way (by encoding each fact separately with LSTM, and attending on each by the decoder as in [31]), but relies completely on the *generation probabilities* without any copy mechanism.

3.4 Training

We train all the models described in this section using the same loss function optimization. More precisely, given a model M that produces a probability $p_t(w|y_{<t})$ of generating token w at decoding step t, we train the whole network end-to-end with the negative log-likelihood loss function of

$$J_{\text{loss}}(\mathbf{x}, \mathbf{y}, \{\mathbf{f}^{(i)}\}_{i=1}^{K}; \mathbf{\Theta}) = -\frac{1}{|\mathbf{y}|} \sum_{t=1}^{|\mathbf{y}|} \log(p_t(y_t|y_{<t}, \mathbf{x}, \{\mathbf{f}^{(i)}\}_{i=1}^{K}))$$

for a training sample $(\mathbf{x}, \mathbf{y}, \{\mathbf{f}^{(i)}\}_{i=1}^{K})$ where $\mathbf{\Theta}$ denotes all the learnable model parameters.

4 Experiments

In this section, we describe the details of dataset, training process, evaluation metrics, and the performance results of baseline and proposed models along with a discussion.

4.1 Dataset

We perform experiments for our problem setup on the recently released CONVAI2 *conversational AI challenge* dataset, which is an extended version of PERSONACHAT [29]. The conversations in CONVAI2 are obtained by asking a pair of crowdworkers to chat with each other naturally based on their randomly assigned personas (from a set of 1155 personas) towards getting to know each other. Personas are created by a different set of crowdworkers, and they consist of ~5 natural language sentences, each describing an aspect of a person that can range from common hobbies like *"I like to play basketball"* to very specific facts like *"I have a pet parrot named Tasha"*, reflecting a wide range of different personalities. The dataset contains ~11000 dialogues with ~160000 utterances, and 2000 dialogues with non-overlapping personas are used for validation and test. For our setting, we use personas as external knowledge sources that models can ground on while generating responses.

4.2 Training and Implementation Details

In all the models explored in this paper, we set the dialogue context to concatenation of the last two dialogue turns separated by a special `CONCAT` token. The models are supplied with the persona facts of the side generating the response at the current turn, while the persona of the other side is concealed. We use a vocabulary of 18650 most frequent tokens and all the remaining tokens are replaced with a special `UNK` token. Embeddings of size 100 are randomly initialized and updated during training. We set the size of LSTM hidden layer to 100 for both encoder and decoder. The encoder and decoder

Model	Perplexity	BLEU	ROUGE-L	CIDEr
MEMNET	61.30	3.07	59.10	10.52
MEMNET + CONTEXTATTENTION	57.37	3.24	59.20	11.79
MEMNET + FACTATTENTION	61.50	2.43	59.34	9.65
MEMNET + FULLATTENTION	59.64	3.26	59.18	12.25
SEQ2SEQ + NOFACT	60.48	3.38	59.46	11.41
SEQ2SEQ + BESTFACTCONTEXT	58.68	3.35	59.13	10.77
SEQ2SEQ + BESTFACTRESPONSE*	49.74	4.02	60.04	16.15
SEQ2SEQ + NOFACT + COPY	58.84	3.25	59.18	11.15
SEQ2SEQ + BESTFACTCONTEXT + COPY	60.25	3.17	59.46	11.17
SEQ2SEQ + BESTFACTRESPONSE + COPY*	38.60	4.54	60.96	21.47
MULTISEQ2SEQ (no COPY)	57.94	2.88	59.10	10.92
DEEPCOPY	**54.58**	**4.09**	**60.30**	**15.76**

Table 1: Results on CONVAI2 dataset. Evaluation metrics on last three columns are better the higher. Perplexity is lower the better. The results of the proposed approach are presented in bold. * indicates that the corresponding model should be considered as a kind of ORACLE because it has access to the fact that is most relevant to the ground-truth response during the inference/test time as defined in Section 3.2.1

vocabularies and embeddings are shared. A shared LSTM encoder is used for encoding both dialogue context and facts of external knowledge source. The model parameters are optimized using Adam [7] with a batch size of 32, a fixed learning rate of 0.001. We apply gradient clipping to 5 when its norm exceeds this value. During inference, we generate responses by employing a beam search of width 4. Our models are implemented in *TensorFlow* [1].

4.3 Results and Discussion

In Table 1, we present our results in comparison with the existing and proposed baseline models. We report the performance of each model across several metrics commonly used for evaluation of text generation models including perplexity, corpus BLEU [14], ROUGE-L [9], CIDEr [24].

As expected, SEQ2SEQ + BESTFACTRESPONSE model and its +COPY version outperform all the other models across all the evaluation metrics. This model pinpoints the importance of selecting the most suitable fact in the persona for the response to be generated at each turn, justifying our underlying motivation for conducting this experiment as highlighted in Section 3.2.1. However, the most suitable fact for the response is not available in the real application scenario, where the models are responsible for picking the useful pieces of information pertaining to the current dialogue turn to generate meaningful responses. Our proposed SEQ2SEQ + BESTFACTCONTEXT model and its +COPY version, on the other hand, are valid baselines for this scenario where the best fact is selected completely based on the dialogue context without relying on the ground-truth response. This model outperforms the previously proposed memory network based model MEMNET [6] for knowledge grounded response generation on all the evaluation metrics, demonstrating its effectiveness despite the fact that it does not have access to all the facts unlike [6]. However, this approach has the following potential weaknesses: (i) if the best persona fact selected w.r.t dialogue context is wrong (irrelevant) for the ground-truth response, the generated response might be drastically misled, and furthermore it is difficult for model to recover from this error because it has no access to other facts, (ii) selecting the best fact w.r.t dialogue context based on *tf-idf* similarity may result in poor fact selection when the lexical overlap between context and response is small which might be a common case especially for the CONVAI2 dataset as the focus of conversation may often change swiftly across the dialogue turns. The latter might be the reason why copying does not help much for this model since it might end up copying irrelevant tokens in the scenario mentioned above.

Our proposed DEEPCOPY model is designed to effectively address the aforementioned issues, where it has access to the entire set of persona facts per dialogue from which it is expected to pick the useful pieces of information while generating the response. DEEPCOPY model outperforms all the models reported in Table 1 except for SEQ2SEQ + BESTCONTEXTRESPONSE models, which we already deem as kind of an upper bound because it has access to the most relevant fact to the response. This justifies the effectiveness of DEEPCOPY model compared to the existing works [6, 29] and the additional baselines we explored in this work. On the other hand, MULTISEQ2SEQ performs considerably worse than DEEPCOPY model despite the fact they both have access to the entire set of

Persona Facts
1. i'm a clean eater. 2. i'm a cancer survivor. 3. my parents were both very athletic. 4. **i got a new job just yesterday to be a life coach.** 5. i love running and preparing for marathons.

Model	Conversation/Response
PERSON1	i really miss it but i have been eating healthy ever since i overcame cancer
PERSON2	omg i am glad you did , do you work now ?
MEMNET*	yes, i do not work, but i do not work.
SEQ2SEQ*	no i ' ve a job at a restaurant
SEQ2SEQ**	i'm a life coach
DEEPCOPY	yes, i just got a new job

Table 2: Example dialogue where the previous two turns from PERSON1 and PERSON2 along with the responses generated by the models acting as PERSON1 are shown on the right. Persona facts for PERSON1 are provided on the left, among which the one in bold is the best fact w.r.t response. MEMNET*, SEQ2SEQ*, SEQ2SEQ** are abbreviations for MEMNET + FULLATTENTION, SEQ2SEQ + BESTFACTRESPONSE, SEQ2SEQ + BESTFACTRESPONSE +COPY models, respectively.

facts and employ the same encoder-decoder architecture except for the copying mechanism. This further justifies the effectiveness of incorporating the proposed hierarchical pointer networks in DEEPCOPY because integrating the external knowledge simply by employing multi-source attention as in [31] does not yield to a good solution with competitive results, performing even worse than SEQ2SEQ + NOFACT on three of the evaluation metrics.

It is also important to note here that DEEPCOPY model, besides helping to generate OOV tokens, may implicitly facilitate learning to attend on better persona facts by increasing the final output probability of tokens that appear in facts. More precisely, if the model assigns high final copy probabilities (computed by the fusion of fact-level and token-level attention) to wrong tokens, then higher loss will be incurred because of wrongly attended facts. Hence, the proposed copy mechanism plays a unique anchor role for reinforcing better fact selection as a side product. This potentially explains the considerable discrepancy between the performance of DEEPCOPY and MULTISEQ2SEQ.

In Table 2, we present an example dialogue where DEEPCOPY model generates a meaningful and fluent response by effectively mixing *copy* and *generate* modes. We can observe that it is able to attend on the right persona fact by taking the dialogue context (especially the question at the end of PERSON2's turn) into consideration. Furthermore, attending to the tokens of this fact, it produces a fluent and valid answer to yes/no question by generating "yes" and copying the rest (and most) of the tokens from the fact. Although it copies most of the tokens from the fact, it is good to observe that it copies exactly the relevant pieces instead of just copying the entire fact. SEQ2SEQ + BESTFACTRESPONSE + COPY model's response is also meaningful and fluent although it may not be as engaging for the continuation of dialog. However, the quality of the response by SEQ2SEQ + BESTFACTRESPONSE quickly degrades compared to its +COPY version. Although the response is still fluent and relevant to the dialogue context, it becomes rather irrelevant to the persona as the model seems to have difficulty of picking the useful information from even the best persona fact it is provided with when the copy mechanism is disabled. Lastly, the response generated by MEMNET+FULLATTENTION model seems to still suffer from repetition, semantic consistency, and relevancy problems that were observed and reported by previous work.

5 Conclusion and Future Work

In this paper, we propose a hierarchical pointer network for knowledge grounded dialogue response generation. Our approach extends the pointer-generator network to enable the decoder to simultaneously copy tokens from the available set of relevant external knowledge in addition to dialogue context. We demonstrate the effectiveness of our approach in comparison with several baselines by experiments on the CONVAI2 dataset. In the future, we plan to apply our model to larger scale datasets of the same fashion where the dialogue is accompanied by a much larger set of knowledge facts (e.g., Wikipedia articles) [5]. This could be done by adding a retrieval component which identifies a few contextually relevant facts [6], which can be used as input to DEEPCOPY. Furthermore, we plan to conduct a human evaluation to better analyze the responses of different models from various aspects such as their appropriateness, coherence with facts, and the extent to which the facts are used.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[4] Sumit Chopra, Michael Auli, and M. Alexander Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *The North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016.

[5] Michel Galley, Chris Brockett, Xiang Gao, Bill Dolan, and Jianfeng Gao. End-to-end conversation modeling: Moving beyond chitchat. `http://workshop.colips.org/dstc7/proposals/DSTC7-MSR_end2end.pdf`, 2018. Online; accessed 23 October 2018.

[6] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[8] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003. Association for Computational Linguistics, 2016.

[9] C.Y. Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

[10] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498. Association for Computational Linguistics, 2018.

[11] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Empirical Methods on Natural Language Processing (EMNLP)*, 2015.

[12] Yishu Miao and Phil Blunsom. Discrete generative models for sentence compression. In *Empirical Methods on Natural Language Processing (EMNLP)*, 2016.

[13] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Computational Natural Language Learning (CoNLL)*, 2016.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[15] Dinesh Raghu, Nikhil Gupta, and Mausam. Hierarchical pointer-generator network for task oriented dialog. *arXiv preprint arXiv:1805.01216*, 2018.

[16] Alan Ritter, Colin Cherry, and William B Dolan. 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, 2011.

[17] M. Alexander Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Empirical Methods on Natural Language Processing (EMNLP)*, 2015.

[18] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[19] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[20] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[21] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Fergus Rob. End-to-end memory networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[22] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[23] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. Abstractive document summarization with a graph-based attentional neural model. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[24] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv:1411.5726*, 2014.

[25] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[26] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[27] Jason Weston, Emily Dinan, and Alexander H. Miller. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776v2*, 2018.

[28] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. In *International Conference on Learning Representations (ICLR)*, 2017.

[29] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[30] Qingyu Zhou, Na Yang, Fur Wei, and Ming Zhou. Selective encoding for abstractive sentence summarization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[31] Barret Zoph and Kevin Knight. Multi-source neural translation. In *The North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016.