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@iee.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. 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 proposed approach extends pointer-generator networks (See et al., 2017) 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 (Ghazvininejad et al., 2018; Zhang et al., 2018) through both automatic evaluation metrics and human evaluation on CONVAI2 dataset.

1 Introduction

Recently, deep neural networks have achieved state-of-the-art results in various tasks including computer vision, natural language and speech processing. Specifically, neural sequence-to-sequence models (Sutskever et al., 2014; Bahdanau et al., 2015) have led to great progress in important downstream NLP tasks like text summarization (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Tan et al., 2017; Yavuz et al., 2018), machine translation (Cho et al., 2014; Sutskever et al., 2014; Luong et al., 2015; Bahdanau et al., 2015), and reading comprehension (Xiong et al., 2017). However, achieving satisfactory performance on dialogue 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 were reported by previous research (Serban et al., 2016; Vinyals and Le, 2015). Most common problems include inconsistency in personality, dull and generic responses, and unawareness of long-term dialogue context. To alleviate these limitations, we turn our focus on a different problem setting for dialogue response generation 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 personality inconsistency issue, we ground our dialogue generation model on external knowledge facts which are a list of persona descriptions in our application (Li et al., 2016a; Zhang et al., 2018). We explicitly use the dialogue history as memory for the model to condition on which potentially encourages a more natural conversation flow. Towards encouraging generation of more specific and appropriate responses while avoiding generic and dull ones, we use a hierarchical pointer network in our model such that it can copy content from two sources: current dialogue history and persona descriptions.

In this work, we propose a novel and general ar-

*Work done while interning at Google AI.

Proceedings of the SIGDial 2019 Conference, pages 122–132
Stockholm, Sweden, 11-13 September 2019.

chitecture 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 (See et al., 2017), 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 (Ghazvininejad et al., 2018; Zhang et al., 2018) 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 (Ritter et al., 2011; Vinyals and Le, 2015). While these studies resulted in a paradigm change compared to earlier work, they do not include mechanisms to represent conversation context. To tackle this problem and have a better representation of conversation context as input to generation, (Serban et al., 2016) 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 often generic and contentless (Vinyals and Le, 2015). 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 (Serban et al., 2017).

Another challenge for dialogue response generation is the integration of knowledge into the generated responses. (Liu et al., 2018) extracted facts relevant to a dialogue from knowledge using string matching, named entity recognition and linking, found additional entities from knowledge that are most relevant to the facts by a neural similarity scorer, and used these as input context features for the dialogue generation RNN. (Ghazvininejad et al., 2018) 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 a baseline.

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). Recently, (Zhang et al., 2018) 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. (Li et al., 2016a) 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 (Raghu et al., 2018) applies hierarchical memory network for task oriented dialog problem. In this work, we compare our model with (Zhang et al., 2018) which uses 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 (Sukhbaatar et al., 2014) and pointer-generator networks (See et al., 2017). Subsequently, we introduce the proposed DEEPCOPY model with a 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 (Weston et al., 2018) 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 (Ghazvininejad et al., 2018; Zhang et al., 2018) and the in-house ones we propose as additional baselines.

3.2.1 Seq2Seq

In a sequence-to-sequence model with attention (Bahdanau et al., 2015), 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 (Ghazvininejad et al., 2018; Zhang et al., 2018) are a more principled approach to incorporating external facts. Similar to (Ghazvininejad et al., 2018), 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 (Ghazvininejad et al., 2018), 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:

MEMNET. This is equivalent to the model used in (Ghazvininejad et al., 2018), described above. This is essentially 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 the *generative profile memory network* (Zhang et al., 2018), where we apply attention only on facts, and we set the decoder's initial state to the combined summary $\hat{u}$.

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 $c_t^{(f)}$ (Zoph and Knight, 2016).

3.2.3 Seq2Seq with Copy Mechanism

Seq2seq models can only generate tokens present in a fixed vocabulary obtained from the training set. Pointer-generator network (See et al., 2017) extends the attentional sequence-to-sequence model (Bahdanau et al., 2015) by employing a pointer network (Vinyals et al., 2015). It has two decoding modes, copying and generating, which are com-

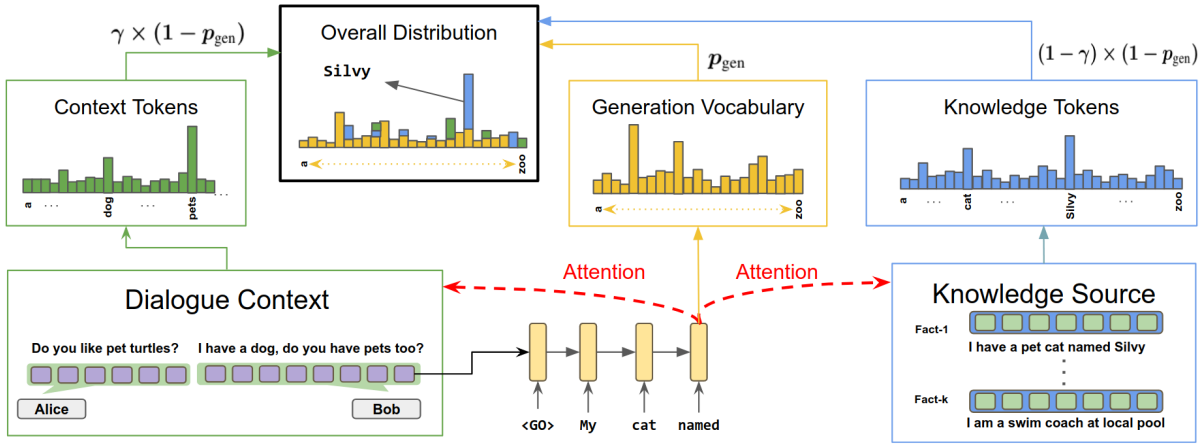

Figure 1: Overview of our proposed approach as described in Section 3.3. The decoder state d_t is used to attend over dialogue context and knowledge source to generate distributions for copying tokens from these sources. The decoder outputs a distribution over a fixed vocabulary. The three distributions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

bined 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 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 (See et al., 2017) can only copy tokens from the encoder input. In this section, we present our proposed DEEPCOPY model that extends pointer-generator network (See et al., 2017) 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 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 $W_2 \in \mathbb{R}^{p \times 2p}$.

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

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

$$c = \sum_{i=1}^{n} \alpha_i u_i \quad (3)$$

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),

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

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

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

Copying from Facts: Hierarchical Pointer Network. We introduce the hierarchical pointer net-

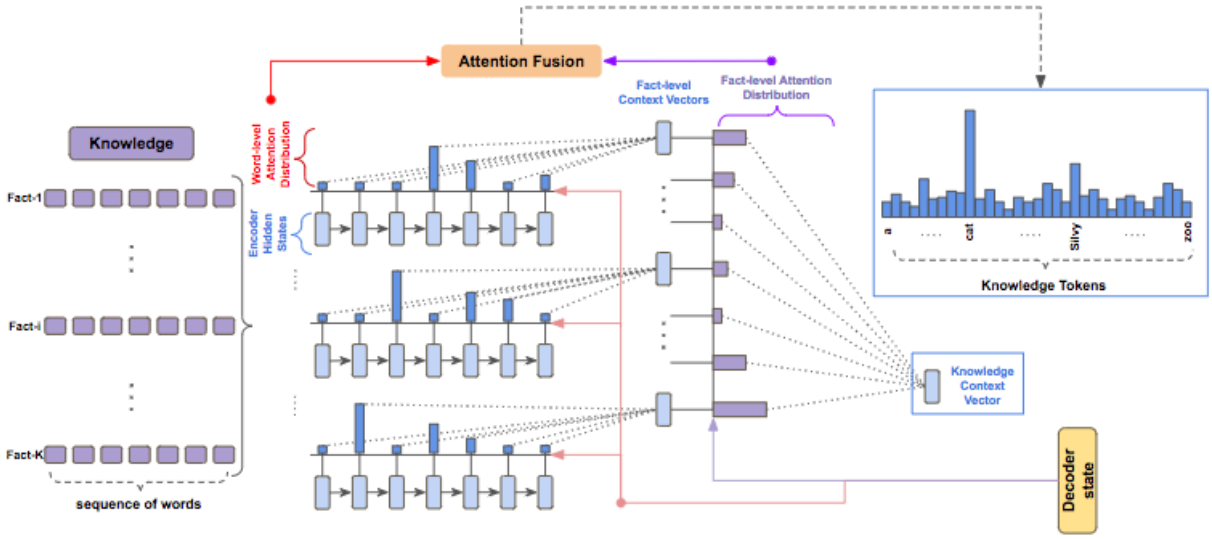

Figure 2: Illustration of hierarchical pointer network. The decoder state d_t is used to attend over tokens for each fact and also over the fact-level context vectors obtained by weighted average of token-level representations (w.r.t token-level attention weights) for each fact. The token-level attention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

work (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 (Equation 8) and the overall fact-level representations of each fact (Equation 9) by

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

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

$$\beta_t, c_t^{(f)} = \texttt{Attention}(\{c_t^{(f)(i)}\}_{i=1}^{K}, h_t) \quad (9)$$

to compute the probability of copying a word w from facts as

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

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.

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

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

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 (See et al., 2017), 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 (Zoph and Knight, 2016)), but relies completely on *generation probabilities* without a 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

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

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 DEEPCOPY model in comparison to proposed and existing baselines.

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 (Zhang et al., 2018). 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 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 (Kingma and Ba, 2015) 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* (Abadi et al., 2016).

4.3 Main Results

In this section, we present the experimental results in terms of both automatic measures and human evaluation.

4.3.1 Automatic Evaluation

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 (Papineni et al., 2002), ROUGE-L (Lin and Och, 2004), CIDEr (Vedantam et al., 2014).

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 (Ghazvininejad et al., 2018) 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 (Ghazvininejad et al., 2018). 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 misinforming, 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

Model	Perplexity	BLEU	ROUGE-L	CIDEr	Appropriateness
[M-1] MEMNET	61.30	3.07	59.10	10.52	3.14 (0.51)
[M-2] MEMNET + CONTEXTATTENTION	57.37	3.24	59.20	11.79	3.41 (0.54)
[M-3] MEMNET + FACTATTENTION	61.50	2.43	59.34	9.65	1.45 (0.25)
[M-4] MEMNET + FULLATTENTION	59.64	3.26	59.18	12.25	3.20 (0.49)
[S2S-1] SEQ2SEQ + NOFACT	60.48	3.38	59.46	11.41	3.12 (0.52)
[S2S-2] SEQ2SEQ + BESTFACTCONTEXT	58.68	3.35	59.13	10.77	3.08 (0.45)
[S2S-3] SEQ2SEQ + BESTFACTRESPONSE*	49.74	4.02	60.04	16.15	2.97 (0.51)
[S2SC-1] SEQ2SEQ + NOFACT + COPY	58.84	3.25	59.18	11.15	3.64 (0.54)
[S2SC-2] SEQ2SEQ + BESTFACTCONTEXT + COPY	60.25	3.17	59.46	11.17	3.60 (0.51)
[S2SC-3] SEQ2SEQ + BESTFACTRESPONSE + COPY*	38.60	4.54	60.96	21.47	3.83 (0.46)
[M-S2S] MULTISEQ2SEQ (no COPY)	57.94	2.88	59.10	10.92	3.32 (0.44)
DEEPCOPY†	**54.58**	**4.09**	**60.30**	**15.76**	**3.67** (0.59)
G.TRUTH	N/A	N/A	N/A	N/A	4.40 (0.45)

Table 1: Main 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. † indicates that the improvement of DEEPCOPY in automatic evaluation metrics over each of the other models (except S2SC-3) is statistically significant with p-value of less than 0.001 on the paired t-test.

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 include the useful pieces of information in 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 (Ghazvininejad et al., 2018; Zhang et al., 2018) and the additional baselines we explored in this work. On the other hand, MULTISEQ2SEQ performs considerably worse than the DEEPCOPY model despite the fact they both have access to the entire set of facts and employ the same encoder-decoder architecture except for the copy 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 (Zoph and Knight, 2016) does not yield to a good solution with competitive results, performing even worse than SEQ2SEQ + NOFACT on 3 of the metrics.

4.3.2 Human Evaluation

Although automatic metrics provide tangible information regarding the performance of the models, we augment them with human evaluations for a more comprehensive analysis of the resulting model generated responses. Towards this end, we randomly sample 100 examples from test data and ask human raters to evaluate the candidate model generated responses in terms of appropriateness. Each example is rated by 3 raters, who are shown a dialog history along with a set of persona facts (of the person in turn), and asked to rate each response based on its *appropriateness* in the dialogue context with a score from 1 (worst) to 5 (best).

In Table 1, we present the results of human evaluation under the *appropriateness* column. Since each response is rated by 3 different human raters, we report the average rating along with the standard deviation in parenthesis. We observe that DEEPCOPY outperforms both the existing memory-network baselines and the proposed sequence-to-sequence baselines on the appropriateness evaluation. It also achieves a performance that is close to the *oracle* model (S2SC-3), which has a leverage of having an access to the fact that is most relevant to the ground-truth response during the inference time. Overall, human evaluation of the responses in terms of appropriateness further justifies the promise and effectiveness of our proposed DEEPCOPY model.

	Diversity	Fact-Inclusion			Agreement
Model	**Distinct**-2 / 3 / 4	**F.Inc**	**F.Per**	**F.Hal**	**F.Inc / F.Per**
M-1	.004 / .006 / .010	0.41	0.01	0.40	0.99 / 0.99
M-2	.010 / .019 / .031	0.43	0.01	0.42	0.97 / 0.99
M-3	.001 / .001 / .002	0.06	0.04	0.02	0.99 / 0.99
M-4	.054 / .010 / .156	0.51	0.09	0.42	0.98 / 0.98
S2S-1	.012 / .022 / .036	N/A	N/A	N/A	N/A / N/A
S2S-2	.012 / .022 / .035	0.54	0.04	0.50	0.97 / 0.99
S2S-3	.026 / .043 / .061	0.79	0.16	0.63	0.97 / 0.97
S2SC-1	.039 / .069 / .104	N/A	N/A	N/A	N/A / N/A
S2SC-2	.035 / .067 / .109	0.73	0.36	0.37	0.99 / 0.99
S2SC-3*	.058 / .111 / .178	0.73	0.55	0.18	0.98 / 0.96
M-S2S	.035 / .065 / .104	0.47	0.05	0.42	0.96 / 0.98
DEEPCOPY	**.059 / .121 / .201**	0.62	0.23	0.39	0.95 / 0.97
G.TRUTH	0.35 / 0.66 / 0.84	0.76	0.49	0.27	0.93 / 0.96

Table 2: Lexical diversity and fact inclusion analysis results. Model names are abbreviated according to Table 1. **F.Inc** denotes the ratio of responses that include factual information. **F.Per** and **F.Hal** denote the ratio of responses where the included fact is consistent with the persona or a hallucinated one, respectively. **Agreement** column corresponds to Cohen's κ statistic measuring inter-rater agreement on binary factual evaluation metrics for **F.Inc** and **F.Per**. * indicates the **ORACLE** model.

Appropriateness scores also demonstrate the advantage of incorporating the soft copy mechanism. Comparing S2S (and M-S2S) models to their copy-equipped counterparts (S2SC) (and DEEPCOPY) in Table 1 immediately reveals a significant gain in appropriateness score. Another significant observation to note here is that ground-truth responses obtain an average appropriateness score of $4.4/5$, which reflects both the noise in CONVAI2 dataset and the difficulty of generating the perfect response even for humans.

4.4 Further Analysis and Discussion

Lexical Diversity Analysis. In Table 2, we report the lexical diversity results using the distinctness metric introduced in (Li et al., 2016b). *distinct-n* score corresponds to the number of distinct n-grams divided by total number of generated n-grams. We can clearly observe that DEEPCOPY generates the most diverse responses among all the models including the copy-augmented oracle model (S2SC-3). Hence, diversity results further show that our proposed model is promising in addressing the most commonly observed *generic response problem* more effectively than existing models by generating more diverse responses.

Fact Inclusion Analysis. We also conduct an analysis on the kinds of factual information included in the model-generated responses. More precisely, our goal is to understand how often the generated response includes a factual information (F.Inc), and whether this information is consistent with the persona facts (F.Per) or a hallucinated one (F.Hal). A good model can naturally include available facts from the persona and hallucinate others when the conversation context requires them. Towards this end, we ask 3 human raters to label responses with 1 (or 0) based on whether a fact is included, and if so, whether this fact is a persona-fact or not.

In Table 2, we present an analysis for the kinds of factual information included in model generated responses. As can be seen from this analysis, models that have a copy mechanism include more facts from the persona than the ones that do not. Another important observation is that the ground-truth responses include facts from persona only in 49% of the times, which indicates that the provided persona facts remain insufficient to cover the complexity of the high entropy open-ended person-to-person conversations.

In Table 2, we present Cohen's κ score for each model and fact analysis metric pair using the scores from 3 raters for each example. We observe for each model and metric pair a κ statistic of greater than 0.9, which indicates a near perfect agreement among raters. Note that the ratio of hallucinated facts (**F.Hal**) is derived directly from human labels for fact inclusion (**F.Inc**) and persona-fact (**F.Per**). That is why, there is no separate labelling process

for hallucinated facts (**F.Hal**). Hence, there is no κ statistic for **F.Hal** in Table 2.

Error Analysis. A deeper analysis of the examples where DEEPCOPY is assigned a worse appropriateness score than the best performing memory-network based baselines (M-2 and M-4) reveals the following further insights: (i) Some of these examples are corresponding to the cases where a generic response (e.g., "I've a dog named radar", one of the frequent generic responses, completely independent of persona facts) is rated much higher (5 to 1) than factual but slightly off (by a single word in this example) responses (e.g., "I have a dog for a living." coming from the persona fact "I walk dogs for a living."), (ii) In another subset of the analyzed examples, DEEPCOPY model generates a response (e.g., "yes, but I want to become a lawyer.") by incorporating a fact that has already been used in the previous turn of the dialog whereas M-2 produces a generic response (e.g., "that's great. do you have any hobbies?", again irrelevant to facts) which is rated higher. (iii) And most of the remaining cases fall into the class of examples where incorporating knowledge facts breaks the conversation flow, which is a crucial observation specific to this dataset that can also be supported by the low persona-fact inclusion ratio (49%) of ground-truth responses.

4.5 Qualitative Observations

In Figure 3, 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.

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

Figure 3: 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.

5 Conclusion and Future Work

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 through various automatic and human evaluations in comparison with several baselines on the CONVAI2 dataset. Furthermore, we conduct diversity, fact inclusion, and error analysis providing further insights into model behaviors. In the future, we plan to apply our model to datasets of the same fashion where the dialogue is accompanied by a much larger set of knowledge facts (e.g., Wikipedia articles) (Galley et al., 2018). This could be done by adding a retrieval component which identifies a few contextually relevant facts (Ghazvininejad et al., 2018) to be used as input to DEEPCOPY.

References

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. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*.

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

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. *arXiv preprint arXiv:1406.1078*.

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

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

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

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. 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.

Jiwei Li, Michel Galley, Jianfeng Brockett, Chris ad Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *The North American Chapter of the Association for Computational Linguistics (NAACL)*.

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

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. 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.

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

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

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

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

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.

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

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

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 Conference on Artificial Intelligence (AAAI)*.

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 Conference on Artificial Intelligence (AAAI)*.

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

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

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

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

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

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

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

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

Semih Yavuz, Chung-Cheng Chiu, Patrick Nguyen, and Yonghui Wu. 2018. CaLcs: Continuously approximating longest common subsequence for sequence level optimization. In *Empirical Methods on Natural Language Processing (EMNLP)*.

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

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