Focused Attention Improves Document-Grounded Generation

Shrimai Prabhumoye[1*]**, Kazuma Hashimoto**[2]**, Yingbo Zhou**[2]**,**
Alan W Black[1]**, Ruslan Salakhutdinov**[1]

[1]Carnegie Mellon University, [2]Salesforce Research

{sprabhum@cs.cmu.edu, k.hashimoto@salesforce.com}

Abstract

Document grounded generation is the task of using the information provided in a document to improve text generation. This work focuses on two different document grounded generation tasks: *Wikipedia Update Generation* task and *Dialogue response generation*. Our work introduces two novel adaptations of large scale pre-trained encoder-decoder models focusing on <mark>building context driven representation of the document and enabling specific attention to the information in the document.</mark> Additionally, we provide a stronger BART baseline for these tasks. Our proposed techniques outperform existing methods on both automated (at least 48% increase in BLEU-4 points) and human evaluation for closeness to reference and relevance to the document. Furthermore, we perform comprehensive manual inspection of the generated output and categorize errors to provide insights into future directions in modeling these tasks.

1 Introduction

Natural language generation (NLG) systems are increasingly expected to be naturalistic, contentful, and situation-aware due to their popularity and pervasiveness in human life (Reiter and Dale, 2000; Mitchell et al., 2014). This is particularly relevant in dialogue systems (Zhang et al., 2018a; Niu and Bansal, 2018), machine translation systems (Mirkin and Meunier, 2015; Rabinovich et al., 2017), story generation (Fan et al., 2018; Yao et al., 2019), and question answering systems (Gatius, 2017; Reddy et al., 2019).

Despite these mainstream applications, NLG systems face the challenges of being bland, devoid of content, generating generic outputs and hallucinating information (Wiseman et al., 2017; Li et al., 2016; Holtzman et al., 2020; Welleck et al., 2020). Grounding the generation in different modalities

like images (Huang et al., 2016; Mostafazadeh et al., 2017; Shuster et al., 2018), videos (Palaskar et al., 2019; Regneri et al., 2013), and structured data (Banik et al., 2013; Gardent et al., 2017) alleviates some of these issues. Generating natural language from schematized or structured data such as database records, slot-value pair, and Wikipedia Infobox has been explored in prior work (Mei et al., 2016; Wen et al., 2015; Lebret et al., 2016). Although useful, these tasks encounter difficulties such as general applicability (databases may not be available for all domains) and are constrained by the available resources (size of the database).

Document grounded generation mitigates these applicability issues by exploiting the vast availability of data in unstructured form (e.g. books, encyclopedias, news articles, and Wikipedia articles). This enhances the applicability of document grounded generation to a wide range of domains with limited (or no) availability of structured data. Hence, recent work has focused on defining new tasks and carving the scope of the problems (Liu et al., 2018; Prabhumoye et al., 2019; Faltings et al., 2020; Zhou et al., 2018; Dinan et al., 2018).

We focus on two different document grounded generation tasks: (1) Wikipedia Update Generation task (Prabhumoye et al., 2019) and (2) Dialogue response generation (Zhou et al., 2018; Dinan et al., 2018). Prior work has studied these two tasks independently and focused on task specific modeling techniques (Zhao et al., 2020a,b; Prabhumoye et al., 2019). <mark>Our work unifies these tasks and formally shows the similarity in them: presence of a context and a document to ground the information in the generation process.</mark>

Our work introduces two novel improvements to the architectures of large scale pre-trained models (Lewis et al., 2019; Raffel et al., 2019): (1) we focus on building context driven representation of the document, <mark>where the context is taken into account while building the representation of the</mark>

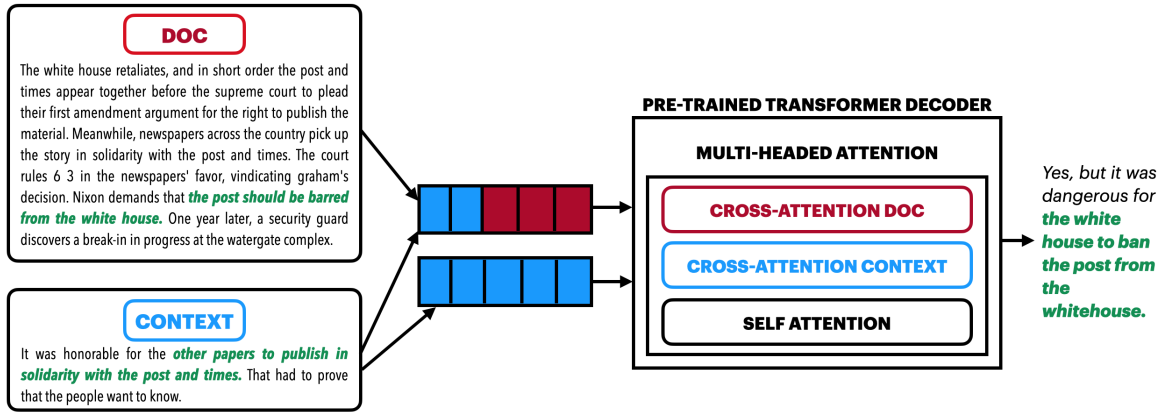---

* Work done during internship at Salesforce.

Figure 1: Document Grounded Generation: An example of a conversation that is grounded in the given document (text in green shows information from the document that was used to generate the response).

document, and (2) during generation we provide specific attention to the information in the document. We provide a stronger BART-based (Lewis et al., 2019) baseline for these tasks. This work shows that pre-trained models albeit good at text generation, can be further improved by providing grounding specific improvements.

Our main contributions are the two new proposed techniques for the document grounded generation tasks (§3.2 and §3.3). We also provide a new baseline which is stronger than the previous state-of-the-art methods (Zhao et al., 2020b; Prabhumoye et al., 2019) for the two tasks. We formally show how the two independent tasks studied in this paper are identical and similar modeling techniques can be used to solve them (§3). Automated and human evaluation results on three different datasets demonstrate substantial improvements (§5.1 and §5.2). Specifically, we achieve an improvement of 19.7 BLEU-4 points compared to Zhao et al. (2020b) on the dialogue generation task. Additionally, significant gains are observed in BLEU-4 compared to BART-based baseline. A comprehensive manual analysis of the generated output is presented in this work which paves way for future work (§6). We will release our code on Github.

2 Task Definition

Our task is to generate text given a context and a source of content (document). Additionally, the generated text should coherently fit the context and contain information from the document. We focus on content present in unstructured form in documents to ground text generation. Figure 1 illustrates such an example. Dialogue response generation is traditionally conditioned on the dialogue context (Vinyals and Le, 2015; Li et al., 2016).

As Figure 1 demonstrates, the generative model is conditioned on both the document as well as the dialogue context. Note that the context and document play different roles in impacting the generation – the context sets the background while the document provides the content necessary to generate the text.

Formally, each sample **i** of our task is defined as a tuple $(\mathbf{d}_i, \mathbf{c}_i, \mathbf{x}_i)$ containing context $\mathbf{c}_i$, document $\mathbf{d}_i$ and text $\mathbf{x}_i$ to be generated. Note that each $\mathbf{d}_i$ can be a single document or a set of documents. The task is to generate $\mathbf{x}_i$ such that it coherently follows $\mathbf{c}_i$ and contains information from $\mathbf{d}_i$. The task can be modeled as the following conditional text generation model: $p_\theta(\mathbf{x}_i|\mathbf{c}_i, \mathbf{d}_i)$, where θ is a set of model parameters.

Figure 1 illustrates that the generator has to account for two inputs the dialogue context $\mathbf{c}_i$ (shown in blue) and the document $\mathbf{d}_i$ (shown in red) to generate the response $\mathbf{x}_i$ grounded in $\mathbf{d}_i$ (text shown in green). If the generative model was only conditioned on dialogue context, then it could produce generic responses like "*Do you think they did the right thing?*" or "*Yes, I agree.*" or hallucinate information like "*Yes, and the Times published it on the front page.*". These which would be appropriate to the given context but are devoid of content or contain wrong information. Document grounded models are capable of responding with interesting facts like "*Yes, but it was dangerous for the white house to ban the post from the white house.*"

3 Methodology

A natural way to model $p_\theta(\mathbf{x}_i|\mathbf{c}_i, \mathbf{d}_i)$ is to train an encoder-decoder model using cross-entropy loss $-\log p_\theta$ with respect to the ground-truth output text. We discuss two ways of building effective representations for encoder-decoder models to focus

on $\mathbf{d}_i$: (1) combine encoder representations of $\mathbf{c}_i$ and $\mathbf{d}_i$, (2) include an additional attention multi-head at each layer of the transformer to specifically focus on the content in $\mathbf{d}_i$.

3.1 Baselines

Low-Res: Zhao et al. (2020a) introduce the state-of-the-art model for document grounded dialogue generation. As described in (§2), the chat history serves as the context $\mathbf{c}_i$ and $\mathbf{x}_i$ is the response to be generated. Zhao et al. (2020a) pre-train their architecture on the dialogue specific Reddit (Dziri et al., 2018) dataset and learn separate parameters for encoding $\mathbf{c}_i$ and $\mathbf{d}_i$. Zhao et al. (2020a) further has three components–context processor, knowledge processor and the language model, each of which build distributions over the vocabulary space. A decoding manager is then trained to generate a token based on these three distributions.

Instead, we employ the recent success of the pre-trained encoder-decoder models (Lewis et al., 2019; Raffel et al., 2019) by using BART (Lewis et al., 2019). One key component of solving this task is to build a representation of the content in the document/s $\mathbf{d}_i$ that is *not* present in the context $\mathbf{c}_i$. We want to leverage the *SelfAttention* feature of transformers (Vaswani et al., 2017) to build such a representation. Since, we use a pre-trained language model as our baseline architecture, we don't use a separate language model component. Instead, we direct our efforts to focus on effectively combining $\mathbf{c}_i$ and $\mathbf{d}_i$.

Content Transfer: Prabhumoye et al. (2019) provide benchmark numbers for the Wikipedia Update Generation task (§2). They explore multiple generative as well as extractive models with and without context. We use their best performing Context Informed LSTM-based encoder-decoder model as baseline. This model concatenates the tokens of the context $\mathbf{c}_i$ and the document $\mathbf{d}_i$ and passes the concatenated sequence to the encoder.

BART: The most straightforward way of using BART for modeling $p_\theta(\mathbf{x}_i | \mathbf{c}_i, \mathbf{d}_i)$ is to concatenate the tokens of the context $\mathbf{c}_i$ and the document $\mathbf{d}_i$ and pass the concatenated sequence ($[\mathbf{c}_i; \mathbf{d}_i]$) to the BART encoder, and then the decoder generates $\mathbf{x}_i$. This is our BART baseline; it already has the advantage of the highly contextualized representations of $\mathbf{c}_i$ and $\mathbf{d}_i$ in comparison with Zhao et al. (2020a). However, fully relying on the self-attention mechanism over the concatenated text would lack the explicit distinction between $\mathbf{c}_i$ and $\mathbf{d}_i$.

Below, we describe two techniques to efficiently build document focused representations. In Figure 1, the method which adds an additional *CrossAttention* multi-head sub-layer to each layer of the transformer is shown. This attention multi-head specifically focuses on the document $\mathbf{d}_i$.

3.2 Context Driven Representation

We propose to use two encoder representations for $\mathbf{c}_i$ and $\mathbf{d}_i$. We first define $\mathbf{h}_d = \mathsf{Encoder}([\mathbf{c}_i; \mathbf{d}_i])$ to get a contextualized representation of $\mathbf{d}_i$, conditioning on the context $\mathbf{c}_i$. $\mathbf{h}_d$ is equivalent to the representation used in the BART baseline. We then apply the same BART encoder to the context alone: $\mathbf{h}_c = \mathsf{Encoder}(\mathbf{c}_i)$. We finally concatenate the encoder outputs $\mathbf{h} = [\mathbf{h}_c; \mathbf{h}_d]$ before passing them to the BART decoder. This $\mathbf{h}$ is **C**ontext **D**riven **R**epresentation (**CoDR**). This method does not require any model architectural modification, and instead the encoder and decoder are fined-tuned to use the multiple input representations.

3.3 Document Headed Attention

In this section, we describe **Do**cument **H**eaded **At**tention (**DoHA**) to further enhance the use of the multiple input representations. A decoder in transformer encoder-decoder models (Vaswani et al., 2017) has two types of multi-head attention mechanism, SelfAttention and CrossAttention with the source sequence. *SelfAttention* module allows each position in the decoder to attend to all positions in the decoder up to and including that position. *CrossAttention* module performs multi-head attention over the output of the encoder stack and attends over the source sequence. While our CoDR method uses the two different source representations, $\mathbf{h}_c$ and $\mathbf{h}_d$, *CrossAttention* is still shared over the concatenated representation $\mathbf{h}$.

In this work, we add an additional multi-head attention *CrossAttention_Doc* to specifically attend over the tokens of the document, while the original *CrossAttention* (named as *CrosstAttention_Cxt*), only attends over the tokens of the context. Each of the multi-heads are of the form:

$$\mathsf{MultiHead}(Q, K, V) = [\mathbf{H}_1; \dots; \mathbf{H}_m]\mathbf{W^o},$$
$$\mathbf{H}_j = \mathsf{Attention}(Q\mathbf{W_j^Q}, K\mathbf{W_j^K}, V\mathbf{W_j^V}).$$

The multi-head function receives three inputs - a query Q, key K and value V. $\mathbf{W^o}$ is an output projection of the concatenated outputs of the attention

4276

heads. Each $\mathbf{H}_j$ is the output of a single attention head and $\mathbf{W}_j^{\mathbf{Q}}$, $\mathbf{W}_j^{\mathbf{K}}$ and $\mathbf{W}_j^{\mathbf{V}}$ are head-specific projections for Q, K, and V, respectively.

Hence, the multi-head *CrossAttention_Doc* is defined by:

$$\mathsf{CrossAttention\_Doc}(Q, K, V)$$
$$= [\mathbf{H}_1; \ldots; \mathbf{H}_m]\mathbf{W}^{\mathbf{do}},$$
$$\mathbf{H}_j = \mathsf{Attention}(Q\mathbf{W}_j^{\mathbf{dQ}}, K\mathbf{W}_j^{\mathbf{dK}}, V\mathbf{W}_j^{\mathbf{dV}}),$$

where $\mathbf{W}^{\mathbf{do}}$, $\mathbf{W}_j^{\mathbf{dQ}}$, $\mathbf{W}_j^{\mathbf{dK}}$ and $\mathbf{W}_j^{\mathbf{dV}}$ are parameters trained specifically to focus on document. The parameters of *CrossAttention_Doc* are initialized with those of *CrossAttention_Cxt*.

Each decoder layer follows the following sequence of functions:

$$\mathbf{h} = \mathcal{F}(\mathsf{SelfAttention}(\mathbf{h_x}, \mathbf{h_x}, \mathbf{h_x})),$$
$$\mathbf{h} = \mathcal{F}(\mathsf{CrossAttention\_Cxt}(\mathbf{h}, \mathbf{h_c}, \mathbf{h_c})),$$
$$\mathbf{h} = \mathcal{F}(\mathsf{CrossAttention\_Doc}(\mathbf{h}, \mathbf{h_d}, \mathbf{h_d})),$$
$$\mathbf{h} = \mathcal{F}(\mathsf{FFN}(\mathbf{h})),$$

where $\mathcal{F}(\mathbf{h})$ is a sequence of $\mathsf{LayerNorm}(\text{residual} + \mathsf{dropout}(\mathbf{h}))$, followed by $\text{residual} = \mathbf{h}$. We integrate the additional attention head *CrossAttention_Doc* by passing the output of the previous attention head *CrossAttention_Cxt* as query. Unlike the weighted attention fusion techniques (Cao et al., 2020), this technique of fusing the additional attention head is novel and useful as it does not require any additional parameters for the fusion.

4 Document Grounded Generation Tasks

Document grounded generation can leverage unstructured data as a source of grounding and can hence be applied to a variety of generation tasks such as dialogue responses, Wikipedia articles, reports and legal argument. This work focuses on *Wikipedia Update Generation* and *Dialogue Response Generation* which have been studied independently in prior work. We discuss the similarities in these two tasks and design a common modeling technique for them.

4.1 Wikipedia Update Generation

This task involves generating an update for Wikipedia context given a news article (Prabhumoye et al., 2019). The dataset was collected by parsing Wikipedia articles and Common Crawl for news articles. It consists tuples of the form $(\mathbf{d}_i, \mathbf{c}_i, \mathbf{x}_i)$, where the grounding document $\mathbf{d}_i$ is

the news article which contains information for the reference update $\mathbf{x}_i$. $\mathbf{x}_i$ is written by a Wikipedia editor as an update to the Wikipedia context $\mathbf{c}_i$. The goal of the task is to generate $\mathbf{x}_i$ given the context $\mathbf{c}_i$ and the document $\mathbf{d}_i$.

4.2 Dialogue Response Generation

Goal oriented dialogues have been traditionally grounded in structured sources like slot-value pairs and databases (Wei et al., 2018; Rastogi et al., 2020). Open domain dialogue generation on the other hand faces the issue of "hallucinating" information (Ghazvininejad et al., 2018). Hence we study open domain dialogue generation which is grounded in documents as a source of information.

CMU_DoG: The CMU Document Grounded Conversations dataset consists of human-human conversations collected over Amazon Mechanical Turk (Zhou et al., 2018). The conversations are grounded in a document provided to the crowdworkers and focuses only on movies. The dataset uses Wikipedia descriptions of movies for grounding the conversations. The dataset consists tuples of the form $(\mathbf{d}_i, \mathbf{c}_i, \mathbf{x}_i)$, where $\mathbf{d}_i$ is a section (or passage) extracted from Wikipedia, $\mathbf{c}_i$ is dialogue history (or context) and $\mathbf{x}_i$ is the reference response. The response $\mathbf{x}_i$ is grounded in $\mathbf{d}_i$ and coherently follows the conversation $\mathbf{c}_i$.

Wizard of Wikipedia: This dataset also consists of human-human conversations collected over Amazon Mechanical Turk and are grounded in passages extracted from Wikipedia (Dinan et al., 2018). These conversations are grounded in a diverse range of topics (totally 1365) which are further split into seen and unseen topics during training and validation. At each step of the dialogue the wizard has access to a set of passages of knowledge which may be relevant to the given dialogue context. The dataset is created by retrieving the top 7 articles (first paragraph only) that are most relevant to the last two turns of dialogue (by wizard and apprentice). Hence, the dataset consists tuples of the form $(\mathbf{d}_i, \mathbf{c}_i, \mathbf{x}_i)$, where $\mathbf{d}_i$ is a list of 7 passages relevant to the conversation, $\mathbf{c}_i$ is dialogue history (or context) and $\mathbf{x}_i$ is the reference response.

The above three tasks consists tuples of the form $(\mathbf{d}_i, \mathbf{c}_i, \mathbf{x}_i)$, where $\mathbf{x}_i$ coherently follows $\mathbf{c}_i$ and is grounded in $\mathbf{d}_i$. Hence, we can use common modeling techniques (§3) for these tasks. [1]

---

[1]Data statistics are shown in Appendix (§A)

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Meteor	F1
Wikipedia Update Generation							
Content Transfer (Prabhumoye et al., 2019)	10.18	4.42	2.20	1.23	10.08	6.21	12.6
BART (baseline)	21.72	14.71	11.28	9.20	22.39	12.90	27.5
CoDR	**25.15**	**17.33**	**13.56**	**11.31**	23.48	**14.38**	29.0
DoHA	25.11	17.04	13.17	10.86	**23.49**	14.28	29.1
CMU_DoG							
Low-Res (Zhao et al., 2020a)	15.00	5.70	2.50	1.20	-	-	10.7
BART (baseline)	23.78	19.27	17.66	16.91	19.30	12.59	21.7
CoDR	26.86	22.75	21.30	20.68	20.41	14.47	22.7
DoHA	**27.33**	**23.05**	**21.55**	**20.90**	**20.44**	**14.55**	22.8
Wizard of Wikipedia (Seen)							
Low-Res (Zhao et al., 2020a)	21.80	11.50	7.50	5.50	-	-	18.0
BART (baseline)	23.92	14.62	10.24	7.75	21.41	15.45	31.1
CoDR	24.00	14.98	10.64	**8.18**	**21.82**	15.71	31.8
DoHA	**24.14**	**15.08**	**10.68**	**8.18**	21.76	**15.89**	31.8
Wizard of Wikipedia (Unseen)							
Low-Res (Zhao et al., 2020a)	20.70	10.10	6.20	4.30	-	-	16.5
BART (baseline)	21.88	12.54	8.44	6.23	19.14	14.03	28.2
CoDR	21.84	12.74	8.60	6.35	19.50	14.22	29.0
DoHA	**22.31**	**13.04**	**8.89**	**6.60**	**19.62**	**14.47**	29.0

Table 1: Results on the automated metrics for the three datasets

5 Experiments and Results

We implement all our models with the transformers tool (Wolf et al., 2019), and the details are in §A.

5.1 Automated Evaluation

Following prior work (Prabhumoye et al., 2019; Zhao et al., 2020a), we evaluate our system-generated sentences against the reference sentences on Rouge-L (Lin, 2004), BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011) metrics.[2] Rouge-L measures the longest common subsequence between the generated sentence and the reference, capturing both lexical selection and word order. METEOR also uses synonyms and stemmed forms of the words in candidate and reference sentences, and thus may be better at quantifying semantic similarities. Additionally, we present F1 which indicates the unigram overlap between the generated output and the reference sentence.[3]

Table 1 shows that the BART baseline outperforms previous state-of-the-art models (Zhao et al., 2020a; Prabhumoye et al., 2019) on all three tasks. It demonstrates that both our improvements DoHA and CoDR perform better than our BART baseline on all metrics and for all three tasks. Notably, we see an improvement of 19.7 BLEU-4 points

on the CMU_DoG dataset compared to Zhao et al. (2020a) which was pre-trained on dialogue specific data; and an improvement on 8.9 BLEU-4 points on the Wikipedia Update Generation compared to (Prabhumoye et al., 2019).[4] We also see substantial improvements (23.6% increase in BLEU-4 for CMU_DoG) compared to the simple BART baseline for the three tasks. In general, DoHA performs slightly better than CoDR on the three tasks.

5.2 Human Evaluation

We follow the human evaluation guidelines mentioned in (Prabhumoye et al., 2019) and evaluate the system generated sentences on three dimensions: (1) closeness of the generated sentences to the references, (2) relevance of the generated sentences to the context and document, and (3) fluency of the generated sentences.

Closeness: The automatic metrics like BLEU, METEOR, and Rouge-L may not be tolerant towards linguistic variations in generated outputs. Hence, we perform a human evaluation to measures how accurately the generated sentence reflects the information in the reference. The annotators are provided with the reference sentence and the generated outputs of two systems labeled A and B in a randomized order. The annotators were instructed to "*Pick the option which is closest in meaning with the reference option.*" The annotators could

---

[2]We use NLG evaluation toolkit (Sharma et al., 2017) from `https://github.com/Maluuba/nlg-eval`

[3]We use the code published at `https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py` to calculate unigram F1.

[4]We use NLG eval script for (Prabhumoye et al., 2019)

Task	BART v CoDR			BART v DoHA			DoHA v CoDR		
	BART	NoPref	CoDR	BART	NoPref	DoHA	DoHA	NoPref	CoDR
	Wikipedia Update Generation								
Closeness	33.3	36.7	30.0	25.5	46.7	27.8	32.2	42.2	25.6
Relevance	18.9	54.4	26.7	24.4	45.6	30.0	33.3	38.9	27.8
	CMU_DoG								
Closeness	15.6	58.8	25.6	30.0	42.2	27.8	33.3	44.5	22.2
Relevance	22.2	43.4	34.4	23.3	42.3	34.4	34.4	42.3	23.3
	Wizard of Wikipedia (seen)								
Closeness	36.7	40.0	23.3	28.9	31.1	40.0	40.5	31.7	27.8
Relevance	24.2	51.6	24.2	32.2	35.6	32.2	28.9	46.7	24.4
	Wizard of Wikipedia (unseen)								
Closeness	23.3	47.8	28.9	44.4	20.0	35.6	21.1	63.3	15.6
Relevance	27.8	47.8	24.4	30.0	43.3	26.6	23.3	41.1	35.6

Table 2: Human evaluation results depicting percentage of times a model was picked (NoPref=No Preference)

select system A or B, or indicate that neither was preferred by picking the third option C. This is a simple evaluation task though potentially biased toward the sole reference.

Relevance: The reference sentence may not be the only correct sentence that fits the context. This is especially true in dialogue generation tasks where contexts like *"How are you?"* and *"What was your favourite part of the movie?"* can have many correct responses that can be produced by grounding on the same document. Hence, we measures whether the generated output contained salient information from the document written in a manner appropriate to the context. The annotators are provided with the document $\mathbf{d}_i$, the context $\mathbf{c}_i$, and the outputs of the two systems A and B, again in a random order. They were instructed to *"Pick the option which contains information from the document and fits the dialogue context coherently"*. Note that the annotators don't have access to the reference in this evaluation. Each judge had to consider whether the information fits with the context and also whether system-generated content could be supported by the document.

Fluency: Finally, we evaluate the fluency of the generated sentences on a scale of 1 (unreadable) to 4 (perfect) as is described in (Zhou et al., 2018).

Human evaluation was conducted on Amazon Mechanical Turk. We conduct 3 comparative studies between the BART, CoDR and DoHA outputs. Each worker was asked to annotated 10 pairs of sentences. We added one control pair among them i.e for $1/10$ pairs, both the sentences were exactly the same. If a worker provides wrong judgement for the control pair then their annotations were

discarded. For each dataset we have total 540 comparative judgements and 90 sentences of each of the models marked for fluency.

Table 2 shows the results of the human evaluation on closeness and relevance. The closeness results show that all the three models BART, CoDR and DoHA generate sentences that are close to the reference, although CoDR and DoHA outperform BART in most cases. Interestingly, the relevance results for Wikipedia Update Generation and CMU_DoG datasets show that CoDR and DoHA generate content that is grounded in the document as opposed to BART. BART baseline generates sentences that are fluent and close to the reference but does not ground in the content of the document as compared to CoDR and DoHA. The 'No Preference' is generally opted over any of the models which is further discussed in §6. For the relevance comparison, annotators have to read a large document to figure out if the generated information is present in the document or not. This can make the annotations noisy especially for Wizard of Wikipedia dataset which has 7 passages as grounding document.

Since both CoDR and DoHA are also BART-based models, the fluency for all three of them is very high and close to each other (BART=3.64, CoDR=3.71, DoHA=3.66).

CoDR and DoHA: The DoHA model still uses the content driven representations ($\mathbf{h}_d$ and $\mathbf{h}_c$). The main difference is that in CoDR model we concatenate $\mathbf{h}_d$ and $\mathbf{h}_c$ and pass it to the decoder but for DoHA we pass $\mathbf{h}_d$ and $\mathbf{h}_c$ separately to the decoder. DoHA has an additional attention layer to focus on the representation of the document $\mathbf{h}_d$

Error Class	%	Chat context	Reference	Generation
Reference and generation are grounded	35	the story is sounding even more interesting. who do you think had the most growth or development in the film?	but there's a twist in that anna has been deceived by the man she was in love with, he was really plotting to take over the kingdom from the sisters the whole time.	but there's a twist in that anna is actually in love with a man who is actually a bad guy. he was actually trying to take over the kingdom from the sisters after elsa injured anna and then their parents died.
Reference is grounded but generation is not	3	hello!	have you heard about the post directed by spielberg?	hello!
Reference is not grounded but generation is	5	yes i agree and how meryl can depict the first female publisher. that would be interesting to see the dynamic of how it effected her career as well	yes, for a woman to have a successful career in those times was not an easy task, for sure.	definitely. i wonder how mcnamara felt when he saw the documents and then told ellsberg that the war was hopeless.
Reference and generation are not grounded	43	the shark was so realistic	yea i live near a beach and it made me not ever want to go swimming	i never thought of it that way, and i agree, it made me terrified to go into the ocean for awhile.

Table 3: Error analysis for the CMU_DoG dataset (only shown for outputs with low Rouge-L score). Text in green indicates grounded information.

only. In this loose sense, DoHA is CoDR plus additional parameters in attention layer to focus on $\mathbf{h}_d$. DoHA performs marginally better than CoDR in automated metrics. But qualitatively (human evaluation) DoHA produces higher quality outputs as compared to CoDR. Table 2 shows DoHA performing better than CoDR on all but one case.

6 Analysis and Discussion

We manually inspect the outputs of the CoDR model on the development set of CMU_DoG and Wikipedia Update Generation dataset to understand the their quality. We inspect 60 samples in each dataset which have Rouge-L score < 60. These are chosen such that we have 10 samples in each of the 6 buckets of Rouge-L score (buckets are range of 10 points: 0-9, 10-19, 20-29, 30-39, 40-49 and 50-59). We analyse the generated outputs along the two aspects of appropriateness of the generation to the context and its grounding in the document.

CMU_DoG: We find that $52/60$ (86.7%) responses were appropriate to the given chat context. These 52 responses are further categorized in Table 3. We found that for about 90% of samples, if the reference is grounded then the generation is also grounded and if the reference is not grounded then the generation is not grounded. Further inspection shows that references are not grounded if they are follow up questions, opinions or experiences that are shared in the conversation. In most of these cases, the context dictates if the response should be grounded or not grounded in the docu-

ment. Since, all of the generated responses in this category are appropriate to the context suggests that these conversational subtleties are not captured by automated evaluation metrics and are given a low score. We also observe a few data artifacts like the mapping of the Wikipedia sections and the chat context is noisy for this dataset. This can be easily resolved by providing all the previous passages of the conversation as grounding to the model. We would also like to note that this dataset was collected under two scenarios: (1) both the people in the conversation have access to the document, and (2) only one person has access to the document. But this distinction is not made in modeling the task. The noise in the dataset can be reduced by modeling only the users that have access to the document in the conversation (similar to Wizard of Wikipedia where only the wizard is modeled).

Wikipedia Update Generation: The error analysis for this task is shown in Table 4. For 5% cases, the reference itself is not grounded in the document. The remaining 95% cases are further classified into 4 error categories. About 85% times, the generation is either completely or partially grounded if the reference is grounded. 43% generations are grounded in document but are linguistic variations of the reference or could be alternate updates to the context. Yet, these are scored low on the Rouge-L metric revealing the inadequacy of the automated metrics. For 23% cases the generation partially hallucinates some information or misses some information present in the reference. 22% times the

Error Class	%	Reference	Generation	R
Linguistic Variation: Reference and generation are grounded and generation is appropriate but a linguistic variation of the reference or an alternate appropriate update.	43	December 12 - The Smiths play Brixton Academy, their last ever gig before their dissolution.	December 12 - The Smiths perform their final show, at Brixton Academy in London.	41
Partial Hallucination: Reference and generation are grounded but generation is either missing or hallucinates some information	23	America Online and Prodigy (online service) offered access to the World Wide Web system for the first time this year, releasing browsers that made it easily accessible to the general public.	The World Wide Web was first introduced on January 17, 1995 on Prodigy.	17
Incoherent Reference: The reference does not coherently follow the context	22	"The Naked Ape", by Desmond Morris, is published.	Zoologist Desmond Morris publishes "The Naked Ape".	26
Incorrect: The generation is either not appropriate or is not grounded (completely hallucinates the information).	7	The year 2000 is sometimes abbreviated as "Y2K" (the "Y" stands for "year", and the "K" stands for "kilo-" which means "thousand").	The Y2K conspiracy theory claimed that a secret nuclear attack by the United States on 2 January 2000 was planned to begin World War 2.	9
Reference is not grounded	5	This was achieved under dead calm conditions as an additional safety measure, whereas the Wrights flew in a 25 mph+ wind to achieve enough airspeed on their early attempts.	This was verified by a video crew present at the test flight.	14

Table 4: Error Analysis for Wikipedia Update Generation task (R denotes Rouge-L score. Text in red indicates hallucinated or missing information.)

reference itself does not seem to coherently fit the context. This is primarily observed for Wikipedia pages that are in the form of a list like *1340s* and *Timeline of DC Comics (1950s)*. Yet, for 50% of the *Incoherent Reference* cases, the generation is grounded in the document and very close to the reference (like the example in Table 4). Only for 7% of the cases, the generation is completely incorrect and hallucinates all of the information. Future work can focus on improving the error in the *Incorrect* and *Partial Hallucination* error classes.

Reference Comparison: With the insights from manual inspection, we performed another comparative study with human judges (on Amazon Mechanical Turk). This was to understand how our models perform in comparison with the reference. The judges are instructed to "*Pick the option that is most appropriate to the given context*". We annotated 100 samples for each DoHA and CoDR model in comparison with the reference on the CMU_DoG and Wikipedia Update Generation datasets. We perform two separate comparative experiments: Reference vs CoDR and Reference vs DoHA. The results in Table 5 show consolidated results for the two models. It shows the total number of times reference was selected, the total number of times 'No Pref' was selected or the total number of CoDR or DoHA was selected. It demonstrates that our mod-

els produce appropriate outputs which can be used as alternate responses/updates. Our models are preferred over the reference in both the tasks suggesting that the automated evaluation is insufficient and the sole reference should not be considered as the only correct response to the context.

7 Related Work

Generation grounded in document has been studied through a large body of summarization work (Rush et al., 2015; Nallapati et al., 2016) and similar tasks such as headline generation (Tan et al., 2017). Multiple new works have extended this research in new directions; Wikipedia Update Generation (Prabhumoye et al., 2019) introduces the task of generating an *update* to the Wikipedia context based on a news document; Wikipedia article generation (Liu et al., 2018) introduces the task of generating an entire Wikipedia article based on multiple documents; Text Editing by Command (Faltings et al., 2020) introduces the task of generating a particular type of Wikipedia edit conditioned on a command provided in natural language and a grounding consisting of snippets of 200 web page results.

Parallely, new tasks have also emerged focusing on document grounding for dialogue response generation (Zhou et al., 2018; Dinan et al., 2018). Zhao et al. (2020a) explore this task in low-resource set-

Dataset	Ref	NoPref	DoHA/CoDR
Wikipedia	33.9	28.3	37.8
CMU_DoG	22.8	45.6	31.6

Table 5: Comparison with reference (Ref) in %age

ting and use pre-training along with a disentangled decoder. The disentangled decoder consists of a context processor, knowledge processor and a language model. A dialogue manager is used to combine the vocabulary distributions provided by these three components. Zhao et al. (2020b) propose a knowledge selection module integrated with pre-trained language models for this task.

Cao et al. (2020) use pre-trained language model GPT-2 (Radford et al.) and explore various attention fusion techniques for persona-based dialogue generation (Zhang et al., 2018b; Dinan et al., 2020). Our DoHA technique also introduces an additional attention multi-head but does not use any additional weights to fuse attention heads. Similarly, Junczys-Dowmunt and Grundkiewicz (2018) use an additional attention multi-head in transformer architecture for automatic post-editing task. We demonstrate how attention can be enhanced in pre-trained models. The CoDR model fuses the representations of the document and the context in the decoder which is inspired by the fusion-in-decoder model in open-domain QA (Izacard and Grave, 2020). Although Bruyn et al. (2020) introduce the usage of BART for knowledge grounded dialogues, it is primarily from the perspective of improving knowledge retrieval. We provide benchmark BART numbers (Table 1) for the generation task. Prabhumoye et al. (2020) provide a schema containing five modules which can be changed to control the generation process. While Zhao et al. (2020a) modify the external input and the output module, we focus on the external input and the generator module of the pre-trained language model.

8 Conclusion and Future Work

This paper proposes two novel improvements for document grounded generation and provides a stronger baseline. This paper demonstrates how similar modeling techniques could be used for two previously separately modeled tasks. Our proposed models outperform the previous techniques and the new stronger baseline on automated metrics and human evaluation for the three datasets discussed in the paper. We present a comprehensive manual inspection which reveal certain data artifacts and

provides us with insight on how to model these tasks in future. Particularly, future work can focus on designing better evaluation metrics which don't penalize linguistic variations in generation. Better models can also be constructed to focus on cases of partial hallucination or incorrect responses.

9 Ethical Considerations

The intended use of the models proposed is to aid the NLG systems in generating content-rich text. Note that this does not imply that the models generate factually correct text. The generation entirely depends on the information in the document provided. If the document itself is factually incorrect then the generation would be grounded in false content and hence generate inaccurate text.

We hope that this technology is used for socially positive applications like building trust of users in dialogue systems like Alexa, Siri and Google Home by providing users with credible information. This work has specifically focused on two datasets of dialogue response generation with the aim that this research not only helps in generating responses which contain useful information but also increase credibility of responses by disclosing the source of information. If dialogue systems base their responses on certain sources of information then they can potentially disclose the source of information to the user. The user then has the agency to make informed decision about trusting the system responses or not.

Additional generations are shown in Appendix (§B). Table 8 and 9 in Appendix §B show the potential misuses of models trained on this task. For both the experiments, a few news articles were hand selected and relevant context was selected from a chosen Wikipedia article. In case of Table 9, the context was curated by hand. Interestingly, the tables also shows the sensitivity of the trained model to the document information. It consists of the same context but different documents were provided as inputs to the model. The generated outputs are different for each document.

References

Eva Banik, Claire Gardent, and Eric Kow. 2013. The KBGen challenge. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 94–97, Sofia, Bulgaria. Association for Computational Linguistics.

M. D. Bruyn, E. Lotfi, Jeska Buhmann, and W. Daelemans. 2020. Bart for knowledge grounded conversations. In *Converse@KDD*.

Yu Cao, Wei Bi, Meng Fang, and Dacheng Tao. 2020. Pretrained language models for dialogue generation with multiple input sources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 909–917, Online. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2018. Augmenting neural response generation with context-aware topical attention. *arXiv preprint arXiv:1811.01063*.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2020. Text editing by command. *arXiv preprint arXiv:2010.12826*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Marta Gatius. 2017. Personalized questions, answers and grammars: Aiding the search for relevant web information. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 203–207, Santiago de Compostela, Spain. Association for Computational Linguistics.

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 *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Hongyuan Mei, TTI UChicago, Mohit Bansal, and Matthew R Walter. 2016. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of NAACL-HLT*, pages 720–730.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Margaret Mitchell, Kathleen McCoy, David McDonald, and Aoife Cahill, editors. 2014. *Proceedings of the 8th International Natural Language Generation Conference (INLG)*. Association for Computational Linguistics, Philadelphia, Pennsylvania, U.S.A.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.

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

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019. Towards content transfer through grounded text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2622–2632, Minneapolis, Minnesota. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, USA.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Engaging image chat: Modeling personality in grounded dialogue. *CoRR*, abs/1811.00945.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: a coarse-to-fine approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4109–4115.

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 Le. 2015. A neural conversational model. In *Proc. ICML Deep Learning Workshop*.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. AirDialogue: An environment for goal-oriented dialogue research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854, Brussels, Belgium. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A   Appendix A

A.1   Implementation Details

We use the transformer toolkit (Wolf et al., 2019) to implement the baseline and both CoDR and DoHA models.[5] Both DoHA (§3.3) and CoDR (§3.2) have the same dimensions and architecture of the BART model (Lewis et al., 2020). For the DoHA model, we initialize *CrossAttention_Doc* with same pre-trained weights of *CrossAttention*. Hence, the layer size of the *CrossAttention_Doc* multi-head is the same as the layer size of *CrossAttention* multi-head in BART. Table 6 shows the maximum sequence lengths used for all the three datasets for both source and target. The data statistics are shown in Table 7.[6] We experimented with two learning rates 5e-5 and 2e-5. We report numbers for the best trained models in each case. Specifically, we report numbers with 5e-5 learning rate for DoHA and CoDR models on the CMU_DoG dataset and the BART baseline for all the three datasets. For Wikipedia Update Generation and Wizard of Wikipedia dataset, we choose the DoHA and CoDR models trained with 2e-5 learning rate. We maintain a common environment (in terms of GPU, operating system, Pytorch version and transformer version) to run all the experiments. We train all the models for 25 epochs.

Zhao et al. (2020a) numbers are directly taken from the paper as the pre-trained model or the generated outputs are not available. We use the same data splits and evaluation toolkits for comparable setting. Hence, Rouge-L and Meteor values are not available for this model. The BLEU, Meteor and Rouge-L numbers are different from (Prabhumoye et al., 2019) due to the usage of different tool-kits in measuring their values.

Dataset	Source Len	Target Len
Wikipedia Update Generation	1024	128
CMU_DoG dataset	512	128
Wizard of Wikipedia	900	40

Table 6: Sequence Lengths

Convergence: Figures 2 and 3 shows the convergence of the baseline BART model in comparison

Dataset	Train	Dev	Test
Wikipedia Update Generation	580.0k	6.0k	50.0k
CMU_DoG	72.9k	4.8k	13.2k
Wizard of Wikipedia	166.7k	17.7k	8.7k

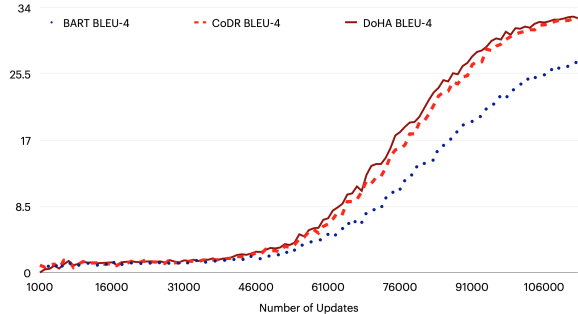Table 7: Dataset Statistics

Figure 2:   Convergence of CMU_DoG development data on the automated metric.

with the CoDR and DoHA models on the development sets of CMU_DoG and Wizard of Wikipedia respectively. We observe that at same number of updates, DoHA and CoDR perform better than BART. This is especially relevant for big datasets like Wikipedia Update Generation which take 15 days to complete 25 epochs.
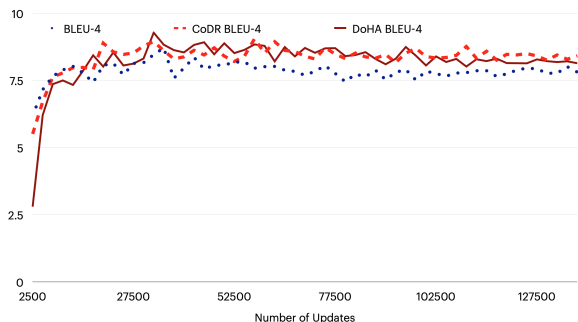

Figure 3:   Convergence of CMU_DoG development data on the automated metric.

B   Appendix B

We hand selected a few news articles which contained inaccurate information. We provided relevant Wikipedia context and generated an update for the context using the trained model on Wikipedia Update generation dataset. The generations are shown in Table 8. Additionally, we provided relevant dialogue context to the model trained on Wizard of Wikipedia dataset (CMU_DoG was not chosen because it is trained only on movie conversations) and the responses generated are shown in Table 9. Note that this shows the potential misuse of the task design.

---

[5]The results are subject to changes in the codebase of the toolkit. Note that we will release our code and trained models to ensure reproducbility of results.

[6]We try to closely follow the processing of the original papers for each of the three datasets.

Document	Context	Generation
`https://www.brad enton.com/latest -news/article237 480339.html`	Climate change includes both the global warming driven by human emissions of greenhouse gases, and the resulting large-scale shifts in weather patterns. Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact on Earth's climate system and the global scale of that impact have been unprecedented.	Though most scientists agree that a "global" climate change is occurring, climate change remains a contentious topic and many scientists, climate scientists, and even climate skeptics believe that it is a hoax.
`https://tinyurl. com/3r5runb`	Climate change includes both the global warming driven by human emissions of greenhouse gases, and the resulting large-scale shifts in weather patterns. Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact on Earth's climate system and the global scale of that impact have been unprecedented.	Some scientists dismiss the existence of climate change.
`https://tinyurl. com/y634czdp`	Climate change includes both the global warming driven by human emissions of greenhouse gases, and the resulting large-scale shifts in weather patterns. Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact on Earth's climate system and the global scale of that impact have been unprecedented.	The Scientific opinion on climate change is that planetwide warming has occurred at least three million years ago.

Table 8: Outputs generated by DoHA model for a few hand selected news articles. The DoHA model is trained on Wikipedia Update Generation dataset.

Document	Context	Generation
`https://tiny url.com/yxed lapz`	What do you think of the rising sea levels?	I think it's crazy that people believe that the oceans will rise and yet people are investing $15 million to live on islands...
`https://tiny url.com/y3u6 fhpo`	Do you think the Affordable Care Act is helping the general public?	It has some great benefits.It has actually resulted in millions of Americans losing their health insurance plan. That is not something former VP Biden claimed.
`https://tiny url.com/3r5r unb`	What do you think of the rising sea levels?	I think the rise of sea levels is a big money making hoax. Why do you think the world is heating up? Any experts?

Table 9: Outputs generated by DoHA model for a few hand selected news articles. The DoHA model is trained on Wizard of Wikipedia dataset.