Controllable Factuality in Document-Grounded Dialog Systems Using a Noisy Channel Model

Nico Daheim\(^1\), David Thulke\(^{2,3}\), Christian Dugast\(^3\), Hermann Ney\(^{2,3}\)
\(^1\)Ubiquitous Knowledge Processing Lab
Department of Computer Science, Technical University of Darmstadt
\(^2\)Chair for Human Language Technology and Pattern Recognition, RWTH Aachen University
\(^3\)AppTek GmbH
www.ukp.tu-darmstadt.de
{thulke, ney}@i6.informatik.rwth-aachen.de

Abstract

In this work, we present a model for document-grounded response generation in dialog that is decomposed into two components according to Bayes’ theorem. One component is a traditional ungrounded response generation model and the other component models the reconstruction of the grounding document based on the dialog context and generated response. We propose different approximate decoding schemes and evaluate our approach on multiple open-domain and task-oriented document-grounded dialog datasets. Our experiments show that the model is more factual in terms of automatic factuality metrics than the baseline model. Furthermore, we outline how introducing scaling factors between the components allows for controlling the tradeoff between factuality and fluency in the model output. Finally, we compare our approach to a recently proposed method to control factuality in grounded dialog, CTRL (Rashkin et al., 2021), and show that both approaches can be combined to achieve additional improvements.

1 Introduction

Recently, document-grounded dialog systems have seen an increase in popularity. Not only are they used to extend task-oriented systems beyond the narrow scope of fixed databases and APIs (Feng et al., 2020; Kim et al., 2020, 2021; Cohen et al., 2022), but also to ground open-domain conversations in information from the web (Zhou et al., 2018; Dinan et al., 2019; Komeili et al., 2022) or in persona descriptions to make dialog agents more interesting and engaging (Zhang et al., 2018). In any of these applications, the conversational system has to use the information from the document and blend it into the response (Roller et al., 2020). This means that the response should remain fluent, grammatically correct and coherent to the previous utterances in the dialog on the one hand, and on the other hand that it correctly reflects the information from the document. This entails that no information is altered and also that no new information should be added if it is not immediately verifiable. This is especially crucial in all cases where a user uses a system to satisfy an information need (Santhanam et al., 2021).

While previous work has shown that retrieving relevant information is a crucial step for task-oriented (Kim et al., 2020) and open-domain document-grounded dialog systems and a potential mitigator of inconsistencies (Shuster et al., 2021), there is sufficient evidence that grounded response generation models may still fail to produce factual responses, even when the correct information is contained in its grounding document. In general, generating outputs that are both fluent and correct remains an open problem not only in dialog systems but natural language generation as a whole (Cao et al., 2018; Maynez et al., 2020; Roller et al., 2021; Ji et al., 2022) which potentially limits its industry adaptation. Models have been found to contradict themselves (Shuster et al., 2022b) or the grounding, and to add additional information that might be harmful and is not verifiable (Shuster et al., 2022a; Ji et al., 2022). Recently, different mitigation strategies were proposed. Cohen et al. (2022) for example use learned discriminators to decide from an n-best list while Rashkin et al. (2021) introduce special control tokens (Keskar et al., 2019) to encourage lexical overlap and entailment between grounding and response. However, discriminators based on estimating human judgments require additional data for training (Cohen et al., 2022) which is costly to obtain. Furthermore, discriminating based on single attributes holds the potential of harming other relevant properties. For
example, discriminating based on factuality can lead to responses mostly repeating their grounding information (Cohen et al., 2022) and might introduce a loss of fluency and dialog coherence. In general, these two goals might be conflicting and their importance depends on the task at hand. While chit-chat lives from engagement and might tolerate inconsistencies, factuality is service-critical in task-oriented systems.

In this work, we present a probabilistic model that inherently combines both of these goals. By factorizing the model according to Bayes’ theorem, we obtain one component that models each goal explicitly. Hence, introducing scaling factors allows for controlling between them. Furthermore, additional unlabeled dialog data can be integrated easily to train one of its components. As directly decoding the model is intractable, we present different approximate decoding schemes for reranking and online decoding that yield significant gains in terms of automatic factuality metrics on several datasets.

2 Related Work

2.1 Document-grounded dialog systems

There has been significant work in document-grounded dialog systems in recent years. A large number of datasets have been proposed for open-domain dialog, in order to facilitate engaging conversations about a variety of topics, such as movies (Zhou et al., 2018), Wikipedia knowledge (Dinan et al., 2019; Dziri et al., 2022a), personal attributes of the agent (Zhang et al., 2018; Dinan et al., 2020) or arbitrary information from the internet (Komeili et al., 2022). Similarly, different task-oriented dialog datasets for information-seeking conversations have been proposed (Kim et al., 2020, 2021; Feng, 2021). Different works have dealt with the problem of document retrieval, for example on batching hard negatives (He et al., 2021) or efficient document retrieval (Thulke et al., 2021), as well as with identifying (Feng, 2021) or rephrasing salient passages within them (Shuster et al., 2022a). Finally, there is also significant work on generating grounded responses using this information, for example in low-resource scenarios (Zhao et al., 2020) or with an emphasis on faithful generations, which we will explore in the following section.

2.2 Hallucination in language generation and dialog

The problem of hallucinations, which one might define as information that is not grounded in the document, dialog context or by common sense, has recently received plenty of attention in neural language generation (Ji et al., 2022), for example in the field of summarization (Cao et al., 2018; Maynez et al., 2020) and dialog systems (Roller et al., 2021). Hence, different types of mitigation strategies that aim to increase the faithfulness of responses have been proposed. Notably, Gabriel et al. (2021) use a set of discriminators to rerank outputs from an n-best list for summarization. For dialog systems, Shuster et al. (2021) show that retrieving relevant information can reduce hallucinations. However, Santhanam et al. (2021) show that correct grounding does not guarantee faithful outputs. Cohen et al. (2022) train discriminators for dialog systems using human judgements, for example to encourage better grounded responses by adding high-quality responses found by the discriminator to the training data. Rashkin et al. (2021) augment the input of a grounded response generation model by additional control tokens (Keskar et al., 2019) to steer generations towards responses entailed by the grounding and Prabhumoye et al. (2021) add an additional attention mechanism to BART that focuses solely on the document.

Along with mitigation strategies, methods for model output and metric evaluation for factuality have been proposed. For example, $Q^2$ (Honovich et al., 2021) proposes a question-answering-based matching and BEGIN (Dziri et al., 2022c) a benchmark for metric evaluation.

Recently, Dziri et al. (2022b) also show that current grounded datasets contain ground-truth responses that further encourage hallucination by being insufficiently grounded and Dziri et al. (2022a) propose FaithDial as a filtered version of Wizard-of-Wikipedia (Dinan et al., 2019) that aims to mitigate this.

2.3 Noisy Channel Modeling in NLP

Given an input sequence $x^T$ and output $y^N$, the noisy channel approach (Shannon, 1948) models the posterior probability of $y^N$ given $x^T$ as $p(y^N | x^T) = p(x^T | y^N)p(y^N) / p(x^T)$. For a long time, such models have been the dominant way of performing Automatic Speech Recognition (ASR) and Machine Translation (MT) (Brown et al., 2020).
In ASR, $p(x^T_1 \mid y^N_1)$ models the acoustic channel (Bahl et al., 1983) and is often called channel model. With the advent of deep learning, discriminative approaches have become popular in both fields and achieve state-of-the-art results (Graves et al., 2006, 2013; Vaswani et al., 2017; Gulati et al., 2020). Nevertheless, the noisy channel approach has recently been explored again for MT (Yu et al., 2017; Yee et al., 2019; Yu et al., 2020; Jean and Cho, 2020; Subramanian et al., 2021), text classification (Min et al., 2022), style transfer (Thulke et al., 2022) and task-oriented dialog systems that are not document-grounded (Liu et al., 2021).

3 Grounded Response Generation

The goal of dialog systems is to find an appropriate system response $u_{T+1}$ conditioned on a sequence of previous turns $u^T_1 := (u_1, \ldots, u_t, \ldots, u_T)$ taken by different interlocutors, where each turn $u_t = [u^N_{t0}]_{N_t} := ([u^N_t][0], \ldots, [u^N_t]_{N_t})$ is a sequence of $N_t$ tokens from the model vocabulary $\mathcal{V}$ prepended with the start of sequence symbol $[u^N_{T+1}]_0 := \langle \text{sos} \rangle$. This is usually done by means of a probabilistic language generation model that models the posterior distribution of the response given the context and is locally-normalized such that the response is generated autoregressively according to

$$p(u_{T+1} \mid u^T_1) = \prod_{n=1}^{N_{T+1}} p([u_{T+1}]_n \mid [u_{T+1}]_{n-1}, u^T_1). \quad (1)$$

In document-grounded dialog systems the response is additionally grounded in information that defines parts of its meaning and is given in the form of unstructured text. For example, in a restaurant booking setting, text on websites might show whether dogs can be brought or in a chit-chat system the grounding information might define the agent’s persona. While in the latter the grounding is known at test-time as a property of the agent, it is usually unknown in the former case. Then, the dialog system has to decide for relevant grounding documents, which are obtained from a document base $D$. In this case, retrieval models that model the distribution $p(d \mid u^T_1, D)$ in order to rank documents $d \in D$ may be employed. By introducing $d$ as a latent variable, the posterior distribution of the response given $u^T_1$ and $D$ is given as follows.

$$p(u_{T+1} \mid u^T_1, D) = \sum_{d \in D} p(u_{T+1}, d \mid u^T_1, D)$$

$$= \sum_{d \in D} [p(d \mid u^T_1, D) \cdot p(u_{T+1} \mid u^T_1, d, D)]$$

$$\approx \sum_{d \in D} [p(d \mid u^T_1, D) \cdot p(u_{T+1} \mid u^T_1, d)]$$

For large $D$ the sum is then approximated by either top-k (Lewis et al., 2020b; Thulke et al., 2021) or maximum approximation. Furthermore, due to the input length restriction in current language generation models, the dependency on $D$ is usually dropped, as outlined in the last step. Using maximum approximation the model becomes

$$\sum_{d \in D} [p(d \mid u^T_1, D) \cdot p(u_{T+1} \mid u^T_1, d)]$$

$$\approx \max_{d \in D} \{p(d \mid u^T_1, D) \cdot p(u_{T+1} \mid u^T_1, d)\}$$

$$\approx p(d \mid u^T_1, D) \cdot p(u_{T+1} \mid u^T_1, d). \quad (2)$$

where $\hat{d}$ is the argument of the maximization of just $p(d \mid u^T_1, D)$ over $D$ since the simultaneous maximization over both components is intractable. This results in a widely-used two-step approach (e.g. (Kim et al., 2020; He et al., 2021)), where retrieval using $p(d \mid u^T_1, D)$ is followed by a response generation model that uses the retrieved document $d$. Then, since the retrieval probability is constant during response generation, we may decide for a response according to the decision rule

$$\hat{d} \mapsto \hat{u}_{T+1} = \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(u_{T+1} \mid u^T_1, d)\}. \quad (2)$$

In line with recent work (Yee et al., 2019; Liu et al., 2021), we will refer to the model used in Equation (2) as direct model in the following. While comparatively simple to use, we note that the model can not use data without grounding annotations directly and has been observed to produce a significant number of incorrect outputs, even with ground-truth grounding (Santhanam et al., 2021; Cohen et al., 2022).

4 A Noisy Channel Approach

We may use the following equivalent decision rule\footnote{A short proof can be found in Appendix A.1. Here, equivalence is meant with respect to the true distributions.}, where the direct model is factorized according to

$$p(x^T_{T+1} \mid \hat{y}^N_{N+1})$$ models the acoustic channel (Bahl et al., 1983) and is often called channel model. With the advent of deep learning, discriminative approaches have become popular in both fields and achieve state-of-the-art results (Graves et al., 2006, 2013; Vaswani et al., 2017; Gulati et al., 2020). Nevertheless, the noisy channel approach has recently been explored again for MT (Yu et al., 2017; Yee et al., 2019; Yu et al., 2020; Jean and Cho, 2020; Subramanian et al., 2021), text classification (Min et al., 2022), style transfer (Thulke et al., 2022) and task-oriented dialog systems that are not document-grounded (Liu et al., 2021).
Bayes’ Theorem:
\[
(u_T^T, d) \mapsto \hat{u}_{T+1} = \arg \max_{u_{T+1} \in V^*} \{ p(u_{T+1} \mid u_T^T, d) \}
= \arg \max_{u_{T+1} \in V^*} \{ p(d \mid u_{T+1}, u_T^T) \cdot p(u_{T+1} \mid u_T^T) \}.
\]

Then, the first component can be formulated as a language generation model
\[
p(d^M \mid u_{T+1}, u_T^T)
= \prod_{m=1}^M p(d_m \mid d_0^{m-1}, u_{T+1}, u_T^T)
\]

The first component can be interpreted as favouring responses that allow to reconstruct the grounding based on the generated response. Hence, one would expect accurate responses to receive higher probability estimates. The second component is an ungrounded response generation model as in Equation (1) and favours fluent responses regardless of the grounding. Furthermore, it can be trained on large amounts of additional dialogues without grounding annotations. In line with previous work on similar models, we will refer to the first component as channel model. Introducing scaling factors between both components then allows for weighting the two objectives of correctness and fluency in order to control the outputs of the model.

We also note that one might arrive at a similar model by the following computation starting from the document-grounded dialog model without maximum approximation:

\[
p(u_{T+1} \mid u_T^T, D)
= \sum_{d \in D} p(d \mid u_T^T, D) \cdot p(u_{T+1} \mid u_T^T, d, D)
= \sum_{d \in D} p(d \mid u_T^T, D) \cdot \frac{p(d \mid u_{T+1}, u_T^T, D) p(u_{T+1} \mid u_T^T, D)}{p(d \mid u_T^T, D)}
= \sum_{d \in D} p(d \mid u_{T+1}, u_T^T, D) \cdot p(u_{T+1} \mid u_T^T, D).
\]

Nevertheless, we leave the exploration of this model to future work.

**Decoding the Noisy Channel Model** Since the channel model has to be evaluated for each hypothesis generated by the response generation model, decoding the model even with beam search is intractable, as \(k \cdot |V|\)-many hypotheses would need to be scored at each iteration with a beam size of \(k\). Therefore, we derive two algorithms to approximately decode the model. First, we introduce reranking, where the noisy channel model is used to score a set of full candidate responses. Then, we introduce online decoding, where the noisy channel model is used to score partial responses during beam search. In both cases, we resort to a proposal model \(q\) to generate candidates. Recall, that we have the following relationship between direct and noisy channel model:

\[
p(u_{T+1} \mid u_T^T, d) = \frac{p(d \mid u_{T+1}, u_T^T) p(u_{T+1} \mid u_T^T)}{p(d \mid u_T^T)}.
\]

During decoding with the maximum-approximated model introduced in the previous section, \(p(d \mid u_T^T)\) is constant and may be dropped. This makes the direct model the natural choice for a proposal model as we would get the same results as with the noisy channel model given the true distributions.

In reranking, the direct model then generates a set of full responses \(U_{T+1}\), from which we decide according to the noisy channel model as follows:

\[
\hat{u}_{T+1} = \arg \max_{u_{T+1} \in U_{T+1}} \left\{ p(d \mid u_{T+1}, u_T^T)^{\lambda_1} \cdot p(u_{T+1} \mid u_T^T)^{\lambda_2} \right\},
\]

where \(\lambda_1 \in \mathbb{R}_{\geq 0}\) and \(\lambda_2 \in \mathbb{R}_{\geq 0}\) are scaling factors. One might also add the direct model without additional computational effort (the probabilities are already calculated during beam search) as an additional factor which has shown beneficial in earlier works (Liu et al., 2021). We note that this resembles the use of discriminators to select responses but does not require additional annotation effort beyond the grounding annotations.

Since reranking is limited by the hypotheses generated by the proposal model, we propose an online decoding algorithm in which the noisy channel model is used during beam search. Since the channel model \(p(d \mid u_{T+1}, u_T^T)\) depends on the final hypothesis that is not available during search, we approximate it using a model \(p(d \mid u_{T+1:n}^T)\) that only depends on partially generated hypotheses similar to Liu et al. (2021). At each step, we score the \(k\) hypotheses in our beam using the noisy channel model. Since scoring all \(k \cdot |V|\) possible extensions is infeasible, we select the best \(k\) extensions only using their direct model score. The noisy channel model score for each of the \(k\) partial hypotheses \([u_{T+1}]_n^n\) up to length \(n\) is then calculated.
Algorithm 1 Pseudocode for Liu et al.’s (−) (Liu et al., 2021) and our (+) decoding algorithm.

**Input** Grounding \( d \), dialog context \( u_1^T \), beam sizes \( k_1, k_2 \)

**Output** Response \( u_{T+1} \)

Beam: \( B = \{ (\text{sos}) \} \)

\[
\text{score}(w) = \log p(w | u_1^T, d) + \lambda_1 \cdot \log p(d | u_1^T) + \lambda_2 \cdot \log p(w | u_1^T)
\]

\[
q(v, w) = \log p(v | u, u_1^T, d)
\]

while end(\( B \)) is False do
  \( B' = \emptyset \)
  for \( w \in B \) do
    \( B' = B' \cup \{ w \circ v \mid v \in \text{top-k}_1 \{ q(v, w) \} \} \)
  end for
  \( B = \text{top-k}_2 \{ \text{score}(w) \} \)
  \( B = \text{top-k}_2 \{ q(v, w) + \text{score}(w) \} \)
end while

\[
\hat{u}_{T+1} = \arg \max_{u_{T+1} \in B} \left\{ \frac{\text{score}(u_{T+1})}{\text{len}(u_{T+1})} \right\}
\]

end

as follows:

\[
p(d | [u_{T+1}]_0^n, u_1^T)^{\lambda_1} \cdot p([u_{T+1}]_0^n | u_1^T)^{\lambda_2}. \quad (8)
\]

Again, we might add the direct model as an additional factor to the score. The algorithm requires locally-normalized models and a channel model trained on partial responses, which we obtain by truncating responses according to a uniform distribution over their length in training.

Lastly, we experiment with the algorithm proposed by Liu et al. (2021) which uses the direct model to generate \( k_1 \) extensions to each of the \( k_2 \) hypotheses in the beam. The \( k_1 \cdot k_2 \) new hypotheses are then pruned back to size \( k_2 \) using the noisy channel model. We outline both algorithms in Algorithm 1.

5 Experiments

We evaluate our model on multiple different English document-grounded dialog datasets by comparing it to the direct modeling objective and the CTRL model (Keskar et al., 2019) presented by Rashkin et al. (2021), where the input is augmented by a sequence of control tokens \( c_{\text{ctrl}}^m \). Thus, the model becomes \( p(u_{T+1} | u_1^T, d, c_{\text{ctrl}}^m) \) and we use the noisy channel model \( p(d | u_{T+1}, u_1^T, c_{\text{ctrl}}^m) \cdot p(u_{T+1} | u_1^T) \) in line with Section 4. We omit the "obligate voice" token, since, for example, Personachat specifically targets conversations where the system responds in first person. The datasets and metrics are described in Section 5.1 and 5.2, respectively. In all our experiments, we finetune the BART-large (Lewis et al., 2020a) checkpoint that is provided as part of the huggingface transformers (Wolf et al., 2020) library, which we further use to implement our experiments. In order to determine the scaling factors for the Noisy Channel model, we do a hyperparameter sweep across \( \lambda_i \in \{0.1, 0.2, \ldots, 2.0\} \) on the validation sets and choose the parameters that perform best in terms of \( Q^2 \). We use \( \lambda_1 = 0.6, \lambda_2 = 0.4 \) for online decoding and \( \lambda_1 = 0.5, \lambda_2 = 0.2 \) for reranking for all experiments after seeing similar trends on all datasets. Furthermore, we always use our proposed online decoding algorithm (see Section 6.3 for a comparison). The results obtained with these experiments are discussed in Section 6.

5.1 Datasets

This section gives a brief overview of the different datasets used in our experiments which capture a variety of settings, for example task-oriented and open-domain dialogs grounded in documents of varying lengths. Dataset statistics can be found in Appendix A.3.

**Personachat** Personachat (Zhang et al., 2018) is a crowdsourced open-domain dialog dataset, where dialogs are grounded in persona descriptions that consist of five short sentences. In our experiments, we use the self configuration and evaluate on the validation split.

**Wizard-of-Wikipedia** Wizard-of-Wikipedia (WoW) (Dinan et al., 2019) is a crowdsourced open-domain dialog dataset, where turns are grounded in Wikipedia articles. Only the wizard can access the grounding in order to teach the apprentice but also choose not to use any grounding. We evaluate on the subset of grounded wizard turns.

**DSTC9** DSTC9 Track 1 is an extension of the MultiWoZ 2.1 dataset (Eric et al., 2020), where turns require information beyond the existing API structure and which was collected from FAQ documents. The test set contains conversations about a new location and a held-out domain, as well as transcripts of spoken conversations.

**Doc2Dial** Doc2dialog (Feng et al., 2020) is a task-oriented dialog dataset, where the agent provides a user with information from public government service websites. The grounding annotations are
Method & Personachat & FaithDial \\
& sBLEU & METEOR & $Q^2$ & BERTScore & F1 & sBLEU & METEOR & $Q^2$ & BERTScore & F1 \\
Direct Model & 5.59 & 17.2 & 0.35 & 0.083 & 14.0 & **15.16** & **41.0** & 0.86 & 0.605 & 69.9 \\
+ Reranking & 5.46 & 17.3 & 0.48 & 0.162 & 22.9 & **14.36** & 40.1 & 0.89 & 0.654 & 74.4 \\
+ Onl Decoding & **5.60** & **19.1** & 0.47 & 0.184 & 25.6 & 13.73 & **39.3** & 0.89 & 0.685 & 77.6 \\
CTRL & 4.87 & 15.6 & 0.63 & 0.256 & 30.6 & 13.65 & 38.0 & 0.92 & 0.725 & 77.8 \\
+ Reranking & 4.14 & 14.4 & 0.70 & 0.303 & 35.7 & 13.19 & 37.9 & **0.93** & **0.749** & 80.6 \\
+ Onl. Decoding & 4.65 & 16.5 & 0.65 & **0.320** & **40.1** & 12.70 & 37.1 & 0.92 & **0.749** & **81.4** \\

Table 1: Main results of our model compared to the direct model and CTRL (Rashkin et al., 2021). We use our online decoding algorithm and all results are within an effective beam size of 30.

given on different levels. We use a concatenation of the annotated spans as a grounding instead of taking entire paragraphs.

**FaithDial** Based on the observation that current dialog datasets contain insufficiently grounded annotations that encourage hallucinations (Dziri et al., 2022b), Dziri et al. (2022a) release FaithDial, an edited version of Wizard-of-Wikipedia that contains significantly less hallucinations. For this, crowdworkers have edited 44% of the grounded training responses from seeker-initiated conversations and all those from the validation and test set.

**5.2 Evaluation Metrics**

In line with the shared tasks on some of the datasets we use for evaluation (Kim et al., 2020, 2021; Feng et al., 2020), we use the sacrebleu (sBLEU) (Post, 2018) implementation of BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005) to assess our model generations with word-overlap based metrics. In addition to that, we use BERTScore (Zhang et al., 2020) and the token-level F1-Score between $u_{T+1}$ and $d$, as well as the recently proposed $Q^2$ metric (Honovich et al., 2021) to evaluate the factual consistency of our models. $Q^2$ is a model-based metric that matches the answers, which are derived from response and grounding, to questions generated based on the response using an NLI model and has shown strong correlations with human judgements in system-level evaluation on WoW.

**5.3 Retrieval**

In addition to experiments that use ground-truth grounding, we also experiment with using the outputs of retrieval models, since usually the grounding is not known at inference time. For retrieval we use two architectures. First, a Bi-Encoder (Bromley et al., 1993), where a dialog and document encoder model map $u^T_1$ and each $d \in D$ to a fixed-size dense vector of the same dimension, respectively. The grounding document is determined by nearest neighbor search, i.e. the decision is made for the document whose vector is closest to the dialog vector in the embedding space. In our experiments, the weights of dialog and document encoder are shared and trained using the Triplet loss criterion.

Secondly, we employ a Cross-Encoder which provides strong performance across a variety of tasks but remains too inefficient in order to be used in practice with large $D$ (Reimers and Gurevych, 2019; Humeau et al., 2020; Karpukhin et al., 2020; Thulke et al., 2021). In the Cross-Encoder, dialog context and document are concatenated as the input to a Transformer model that subsequently performs relevance classification such that the document with the highest score is retrieved.

We use RoBERTa-large (Liu et al., 2019) for all experiments use Recall@1 (R@1) for evaluation.
6 Results

Table 1 shows the results obtained with our model using our proposed online decoding algorithm, reranking, and no additional data in training but the corresponding training set. We identify the following trends:

1) Our model consistently outperforms the direct modeling objective in terms of all automated factuality metrics.

2) There is no clear trend in terms of word-overlap-based metrics, where our model and the direct model show comparable performance.

3) CTRL gives larger improvements in terms of $Q^2$ on all datasets but DSTC9. Nevertheless, the additional control tokens may be seen as a data filtering method that adds new information to the training data that is not available to our model.

4) Adding control tokens to our model, i.e. combining CTRL with the noisy channel approach, gives further improvements in terms of factuality metrics.

5) The improvements obtained on the unseen set of Wizard-of-Wikipedia indicate that the model is also able to generalize appropriately to new information.

Overall, the results indicate that by scaling the channel model contribution appropriately, the faithfulness of responses can indeed be improved in comparison to the direct model. In the following, we present further results to understand how the model behaves under different scaling factors (Section 6.1), uncertain retrieval (Section 6.2), different compute budgets (Section 6.3), and the presence of additional data (Section 6.4) before concluding the section with a qualitative analysis (Section 6.5).

6.1 Controllability

Table 2 shows results obtained with noisy channel online decoding with different scaling factors. In addition to the previously mentioned metrics, we report Perplexity as a proxy for fluency (Dinan et al., 2020) and the longest common subsequence.

| $\lambda_1/\lambda_2$ | sBLEU ↑ | PPL ↓ | $Q^2$ ↑ | BERTScore ↑ | LCS |
|----------------------|--------|-------|--------|-------------|-----|
| Direct Model         | 4.89   | 17.2  | 0.64   | 0.27        | 0.66|
| 0.6/0.4              | 4.70   | 19.5  | 0.64   | 0.52        | 0.62|
| Gold response        | 100.0  | 18.3  | 0.26   | 0.09        | 0.25|
| Gold document        | 1.97   | 47.8  | 99.7   | 1.0         | 1.0 |

Table 2: $Q^2$ and Perplexity by ratio of factors on Personachat with beam size 10.
Table 3: Results on the outputs of a Bi-Encoder and Cross-Encoder retrieval model on DSTC9 test.

Table 4: Results with noisy channel model and additional response generation model training data.

Larger beam size improves the performance of the noisy channel model but not the direct model.

6.4 Additional data

In order to study the effect of additional training data for the response generation model, we train the component on all target datasets as well as MultiWoZ 2.1 (Eric et al., 2020) 2, Taskmaster-1 (Byrne et al., 2019), TopicalChat (Gopalakrishnan et al., 2019) and CMU DoG (Zhou et al., 2018). However, as shown in Table 4, we do not see consistent but often dataset-specific improvements. For example, \(Q^2\) tends to be better with more data in reranking but not online decoding, which also did not change in our experiments with different scaling factors.

6.5 Qualitative Analysis

When comparing the generation outputs of all models, we can make the following observations:

1) the outputs of both the noisy channel model and CTRL appear more faithful to the grounding and more specific. For example, on DSTC9 and doc2dial the direct model sometimes leaves out important details, such as that an ID card needs to be shown when someone picks up a ticket at a train station, which both noisy channel model and CTRL incorporate into the response. Furthermore, the direct model appears to generate generic responses more often, such as "do you have a pet?" when the topic in Personachat is "dog" or "cat".

2) In general, qualitative analysis supports the results from our automatic evaluation that a higher channel model factor implies more from the grounding being copied into the response, which however can come at a loss of fluency and coherence.

3) A higher response generation model factor leads to more abstractiveness and a better connected response that, for example, contains follow-up questions more often. On the other hand, a too high factor also led to hallucinations in our experiments.

\(^2\)We remove the dialogs contained in DSTC9.
4) When comparing CTRL and the noisy channel model, the main difference appears to be that the responses of the noisy channel model (with a suitable factor) seem more connected to the previous turns than in CTRL, especially on a chit-chat task like Personachat. Nevertheless, with a higher channel model factor the generations become more similar to those of CTRL. The combination of CTRL and the noisy channel model might be especially suited for task-oriented dialog, where faithfulness is crucial, whereas for some open-domain settings we think that our model may be more suitable without control tokens.

Finally, some example outputs can be found in the Appendix.

7 Conclusion

In this paper, we present a model for response generation in document-grounded dialog that explicitly optimizes for faithfulness and fluency. The model decomposes the posterior distribution of response given context and grounding into two components according to Bayes’ Theorem and thus, by introducing scaling factors, allows for encouraging more correct or more fluent responses. Since decoding the model directly is intractable, we derive and compare different approximate decoding schemes that use reranking or online decoding. We compare the model to directly modeling the posterior distribution of response given context and grounding and a variant of CTRL that was proposed recently to encourage faithfulness in grounded response generation, which we again factorize into two components. An evaluation on five different open-domain and task-oriented dialog datasets shows improvements in terms of factuality on top of both models. Furthermore, we highlight how the scaling factors can be used to control how much the model copies from the grounding and how much weight is given to a well-connected response, and we investigate the influence of additional training data for one of the model components. Lastly, we show that our model also gives improvements when dealing with uncertain document retrieval.

7.1 Future Work

In future work we would like to explore the noisy channel model for document-grounded response generation without maximum approximation, such as the one that we present in Equation (5).

8 Limitations

The main limitation of our model is that the approximate decoding schemes introduce significant computational overhead in comparison to the direct model and CTRL, which amounts to a factor of up to 10 for online decoding with a large beam size using a not yet fully optimized implementation. Furthermore, decoding is a lot more complex and the scaling factors need to be tuned which implies significant additional computations that are necessary and results in a larger carbon footprint. In addition, our noisy channel model has a significantly higher number of parameters than the baseline model. Finally, we mainly rely on automatic metrics to assess the faithfulness of the proposed approaches and leave out a broader evaluation of the general quality of generated responses but for a small-scale qualitative study cf. Section 6.5.

9 Broader Impact

In general, generative dialog systems are a promising field of research and can be less restricted in the topics they can deal with than rule-based models or approaches that use predefined dialog flows, for example, which require a lot of hand-crafting and possibly experts to write appropriate responses. Nevertheless, such language generation-based approaches bear the danger of repeating harmful content and biases that may have been present in the training data or of generating inappropriate responses, in general. Furthermore, some applications require faithful responses by law or are at least critical to the service. While our model can improve the faithfulness of responses, it can not be guaranteed. Therefore, we would not recommend to use the model in these applications. While document-grounded dialog systems in general can be used to ground system responses in helpful and correct real-world information, these systems could potentially also be misused to ground dialog systems in misinformation or other harmful documents.

Acknowledgements

This work has been financially supported by multiple sources: The project HYKIST funded by the German Federal Ministry of Health on the basis of a decision of the German Federal Parliament (Bundestag) under funding ID ZMVI1-2520DAT04A.

The project NeuroSys which, as part of the initiative “Clusters4Future”, is funded by the Fed-
eral Ministry of Education and Research BMBF (03ZU1106DA).

The German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

The European Union under the Horizon Europe grant No. 101070351 (SERMAS).

References

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5(2):179–190.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature Verification Using a Siamese Time Delay Neural Network. In Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993], pages 737–744. Morgan Kaufmann.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2):263–311.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the Original: Fact Aware Neural Abstractive Summarization. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 4784–4791. AAAI Press.

Nico Daheim, David Thulke, Christian Dagust, and Hermann Ney. 2021. Cascaded span extraction and response generation for document-grounded dialog. In Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021), pages 57–62, Online. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhunoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burlsev, and Jason Weston. 2020. The second conversational intelligence challenge (convai2). In The NeurIPS ’18 Competition, pages 187–208, Cham. Springer International Publishing.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In Proceedings of the International Conference on Learning Representations (ICLR).

Nouha Dziri, Ehsan Kamalpoolo, Sivan Milton, Osmar Zaiane, Mo Yu, Edordao Ponti, and Siva Reddy. 2022a. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. ArXiv preprint, abs/2204.10757.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022c. Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark. Transactions of the Association for Computational Linguistics, 10:1066–1083.
Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tür. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 422–428, Marseille, France. European Language Resources Association.

Song Feng. 2021. DialDoc 2021 Shared Task: Goal-Oriented Document-grounded Dialogue Modeling. In Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021), pages 1–7, Online. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8118–8128, Online. Association for Computational Linguistics.

Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021. Discourse Understanding and Factual Consistency in Abstractive Summarization. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 435–447, Online. Association for Computational Linguistics.

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

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006, volume 148 of ACM International Conference Proceeding Series, pages 369–376. ACM.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 6645–6649.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020, pages 5036–5040. ISCA.

Huang He, Hua Lu, Siqi Bao, Fan Wang, Hua Wu, Zhengyu Niu, and Haifeng Wang. 2021. Learning to Select External Knowledge with Multi-Scale Negative Sampling. ArXiv preprint, abs/2102.02096.

Or Honovich, Leshem Choshen, Roeee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. $Q^2$: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In International Conference on Learning Representations.

Sébastien Jean and Kyunghyun Cho. 2020. Log-Linear Reformulation of the Noisy Channel Model for Document-Level Neural Machine Translation. In Proceedings of the Fourth Workshop on Structured Prediction for NLP, pages 95–101, Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. ArXiv preprint, abs/2202.03629.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Nitiish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. ArXiv preprint, abs/1909.05858.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tür. 2020. Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access. In Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papan gelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. “how robust r u?”: Evaluating task-oriented dialogue systems on spoken conversations. In ASRU 2021.
Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Pikus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktaeschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Qi Liu, Lei Yu, Laura Rimell, and Phil Blunsom. 2021. Pretraining the Noisy Channel Model for Task-Oriented Dialogue. Transactions of the Association for Computational Linguistics, 9:657–674.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv preprint, abs/1907.11692.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5316–5330, Dublin, Ireland. 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 of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021. Focused Attention Improves Document-Grounded Generation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4274–4287, Online. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 704–718, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. 2020. Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 300–325, Online. Association for Computational Linguistics.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. In EMNLP 2021 Workshop on NLP for Conversational AI.

C. E. Shannon. 1948. A mathematical theory of communication. The Bell System Technical Journal, 27(3):379–423.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022a. Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion.
Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022b. Am I me or you? state-of-the-art dialogue models cannot maintain an identity. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 2367–2387, Seattle, United States. Association for Computational Linguistics.

Sandeep Subramanian, Oleksii Hrinchuk, Virginia Adams, and Oleksii Kuchaiev. 2021. NVIDIA NeMo’s neural machine translation systems for English-German and English-Russian news and biomedical tasks at WMT21. In Proceedings of the Sixth Conference on Machine Translation, pages 197–204, Online. Association for Computational Linguistics.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient Retrieval Augmented Generation from Unstructured Knowledge for Task-Oriented Dialog. In Workshop on DSTC9, AAAI.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2022. Adapting document-grounded dialog systems to spoken conversations using data augmentation and a noisy channel model. In Workshop on DSTC10, AAAI.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Cunwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and Effective Noisy Channel Modeling for Neural Machine Translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5696–5701, Hong Kong, China. Association for Computational Linguistics.
A Appendix

A.1 Proof of Equivalence of direct and noisy channel model decoding

Given a dialog context $u_T^1$ and grounding $d$, we may define the search problem as finding an appropriate response $u_{T+1}$. Following our derivation from Section 3 we can formalize this according to the decision rule

$$ (u_T^1, d) \mapsto \hat{u}_{T+1} = \arg \max_{u_{T+1} \in V^*} \{ p(u_{T+1} \mid u_T^1, d) \}, $$

(9)

where $V^*$ denotes the set of finite strings that may be constructed from a fixed vocabulary $V$ using the Kleene closure $\ast$. Then, we may use the following equivalent decision rule

$$ (u_T^1, d) \mapsto \hat{u}_{T+1} = \arg \max_{u_{T+1} \in V^*} \left\{ p(d \mid u_{T+1}, u_T^1) \cdot p(u_{T+1} \mid u_T^1) \right\}. $$

(10)

Proof.

$$ \arg \max_{u_{T+1} \in V^*} \{ p(u_{T+1} \mid u_T^1, d) \} $$

$$ = \arg \max_{u_{T+1} \in V^*} \{ p(u_{T+1} \mid u_T^1, d) \cdot p(d \mid u_T^1) \} $$

$$ = \arg \max_{u_{T+1} \in V^*} \{ p(u_{T+1} \mid u_T^1) \} $$

$$ = \arg \max_{u_{T+1} \in V^*} \{ p(d \mid u_{T+1}, u_T^1) \cdot p(u_{T+1} \mid u_T^1) \} $$

Here, the first step is obtained by multiplying with $p(d \mid u_T^1)$, which does not change the argument of the maximization.

A.2 Experiment details

For all of our experiments and models, we use BART-large (Lewis et al., 2020a), which adheres to the standard Transformer model from Vaswani et al. (2017) and consists of 12 encoder and 12 decoder layers with a hidden size of 1024 and 406M parameters. We finetune each model for 10 epochs using an initial learning rate of $6.25 e^{-5}$, with no warmup steps, and linear learning rate decay. We use a batch size of 32 by using gradient accumulation. The model is evaluated on the validation set after each epoch and the model with the smallest eval loss is picked as our final model. We truncate the dialog history at 384 tokens and restrict the length of the grounding to 128 tokens, after which it is cut off.

Except for the Bi-Encoder, which we train using the Triplet Loss, all models are trained using the Cross-Entropy criterion.

All of our experiments using online decoding use $\lambda_1 = 0.6, \lambda_2 = 0.4$ and all of the reranking experiments use $\lambda_1 = 0.5, \lambda_2 = 0.2$.

All models were trained and evaluated on NVIDIA 1080 or 2080 GPU’s.

A.3 Dataset statistics

| Dataset          | Split     | Domain            | # Dialogs | #Documents |
|------------------|-----------|-------------------|-----------|------------|
| DSTC9            | train     | Task-oriented     | 19,184    | 2,900      |
|                  | test      |                   | 1,981     | 12,039     |
| (Kim et al., 2020)|          |                   |           |            |
| Personalchat     | train     | Open-domain       | 10,907    | 1,155      |
|                  | test      |                   | 1,000     |            |
| (Zhang et al., 2018)|        |                   |           |            |
| Wizard-of-Wikipedia| train  | Open-domain       | 18,430    | 93M sentences |
|                  | test_seen |                   | 965       | 93M sentences |
|                  | test_unseen|                   | 968       | 93M sentences |
| (Dinan et al., 2019)|        |                   |           |            |
| FaithDial        | train     | Open-domain       | 4,094     | 93M sentences |
|                  | test      |                   | 791       | 93M sentences |
| (Dziri et al., 2022a)|       |                   |           |            |
| Doc2Dial         | train     | Task-oriented     | 3,474     | 487        |
|                  | test      |                   | 787       | 572        |
| (Feng et al., 2020)|        |                   |           |            |

A.4 Sample model outputs

In this section we present a number of cherry-picked examples for multiple datasets to supplement the qualitative analysis from Section 6.5.

Furthermore, we present examples obtained with different scaling factors to highlight the behavior of the model under these conditions.
| Grounding | i enjoy listening to classical music . i'm a christian . my favorite color is red . i can drive a tractor . my sister is a pole dancer . my mom passed away when i was young . i work in the insurance business . my german shepherd is barnaby . i love playing softball . my father is in the army . i recently was fired and had to get a new job . my mom was a pole dancer . she's very high maintenance . my favorite color is red . i can drive a tractor and she's very high maintenance . |
| Direct Model | what do you do for a living? do you have any pets? what do you do for a living? do you have any pets? do you have any pets? do you have any pets? do you have any pets? do you have any pets? |
| Noisy Channel | CTRL Online | Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? |
| CTRL Online | Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? |
| Grounding | i enjoy listening to classical music . i'm a christian . my favorite color is red . i can drive a tractor . my sister is a pole dancer . my mom passed away when i was young . i work in the insurance business . my german shepherd is barnaby . i love playing softball . my father is in the army . i recently was fired and had to get a new job . my mom was a pole dancer . she's very high maintenance . my favorite color is red . i can drive a tractor and she's very high maintenance . |
| Direct Model | what do you do for a living? do you have any pets? what do you do for a living? do you have any pets? do you have any pets? do you have any pets? do you have any pets? do you have any pets? |
| Noisy Channel | CTRL Online | Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? |
| CTRL Online | Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with? |
**WoW**

| Grounding | She was the founder and namesake of the Chanel brand. |
|-----------|----------------------------------------------------|
| Direct Model | Have you ever heard of Coco Chanel? She was the founder of the Chanel brand. |
| Reranking: | Have you heard of Coco Chanel? She was the founder and namesake of the Chanel brand. |
| Online Decoding: | I love Coco Chanel. She was the founder and namesake of the Chanel brand. |
| CTRL | Coco Chanel was the founder and namesake of the Chanel brand. |
| CTRL Online | Coco Chanel was the founder and namesake of the Chanel brand. |

**Doc2Dial**

| Grounding | Financial difficulties |
|-----------|------------------------|
| Direct Model | Do you have financial difficulties? |
| Reranking: | You would have to have financial difficulties |
| Online Decoding: | Financial difficulties are one of the requirements to get a general forbearance |
| CTRL | Financial difficulties |
| CTRL Online | Financial difficulties |

**Grounding**

| Grounding | Do you have financial difficulties? |
|-----------|------------------------------------|
| Direct Model | You may request a hearing online. |
| Reranking: | Yes, I disagree with the determination you made at the reconsideration level. |
| Online Decoding: | You may then request a hearing online. |
| CTRL | You may request a hearing online. |
| CTRL Online | You may request a hearing online. |

**Direct Model**

| Ground truth | Yes, I disagree with the determination you made at the reconsideration level. |
|-------------|-------------------------------------------------|
| Ground truth | You may then request a hearing online. |
| Reranking: | You can request a hearing online if you disagree with the determination. |
| Online Decoding: | In that case you may request a hearing online. |
| CTRL | You may request a hearing online. |
| CTRL Online | You may request a hearing online. |
| Grounding | In computing, React (sometimes styled React.js or ReactJS) is a JavaScript library for building user interfaces. I’m not sure I’ve heard of it before, what is it? |
|---|---|
| $u_T$ | 0.5 |
| $\gamma$ | 2.0 |
| It is a JavaScript library for building user interfaces. React (sometimes styled React.js or ReactJS) is a JavaScript library for building user interfaces. |

| Grounding | David Suchet stars as the eponymous detective, Agatha Christie’s fictional Hercule Poirot. No do you know what’s it about? |
|---|---|
| $u_T$ | 0.5 |
| $\gamma$ | 2.0 |
| It’s about a detective named Hercule Poirot, played by David Suchet. |

| Grounding | my friends are all skateboarders. i’ve a broken arm. i am a skateboarder. my favourite skater is rob dryden. oh no! what is wrong with it? are you a pro skateboarder |
|---|---|
| $u_T$ | 0.5 |
| $\gamma$ | 1.4 |
| i broke my arm trying to skateboard with my friend rob dryden |

| Grounding | i like to work on vintage cars. i own two vintage mustangs. i like to go to country concerts on weekends. my favorite music is country. i have two dogs. that’s amazing. i’m currently studying spanish. i love to go to country concerts on the weekends. i love to go to country concerts on the weekends with my dogs. |
|---|---|
| $u_T$ | 0.5 |
| $\gamma$ | 1.4 |
| i try to work and stay busy. i am in between jobs right now. |

| Grounding | i don’t drink soda. i’ve a boyfriend. i live in a house. i like to eat hamburgers and french fries. i try to work and stay busy. i am in between jobs right now. what do you do for a living? do you like hamburgers and french fries? i love them. |
|---|---|
| $u_T$ | 0.5 |
| $\gamma$ | 1.4 |
| i volunteer my time with a local bunny rescue. my favorite color is navy blue. i work fulltime in a shipyard. i donate a part of my salary to charity each month. wow that’s exciting good for you. i’m learning about computers, programming. that’s cool. i work at a shipyard. i love computers and programming. cool. i work in a shipyard full time. what do you do in your spare time? |

| Grounding | i’m a descendant of christopher columbus. i love to cook paella. i’ve a weakness for fish and chips. europe must be very beautiful my sister went there. i visit europe twice a year. i am an art major in college. it is. i love it. i’ve been there twice a year since i was in college. i’m from a small town. i was raised on a horse farm. i want to be a lawyer. cool. i work at a shipyard. i love computers and programming. it is. i love it. i go twice a year, and i cook paella and fish and chips. |
|---|---|
| $u_T$ | 0.5 |
| $\gamma$ | 1.4 |
| i volunteer my time with a local bunny rescue. my favorite color is navy blue. i work fulltime in a shipyard. i donate a part of my salary to charity each month. wow that’s exciting good for you. i’m learning about computers, programming. that’s cool. i work at a shipyard. i love computers and programming. cool. i work in a shipyard full time. what do you do in your spare time? |

| Grounding | my family has raised horses for a generations. i dream of moving to the city. i’m from a small town. i was raised on a horse farm. i want to be a lawyer. keep it in your family, that’s nice for future generations. thank you. i want to move to the city to live on a horse farm. thank you. i want to move to the city one day, but i live in a small town. |
|---|---|
| $u_T$ | 0.5 |
| $\gamma$ | 1.4 |
| Opening hours are from 11am to 5pm on Wednesdays only. The museum opens at 11am on Wednesdays and at 5pm on Sundays. Do you have any other questions about the museum? The museum opens at 11am on Wednesdays only. Is there anything else I can help you with? |

| Grounding | Cathedral of Saint Mary of the Assumption is open from Monday to Friday from 9:30 am to 4:00 pm. On Saturday it is open from 11:00 am to 5:30 pm and on Sunday it is open from 9:00 am to 3:30 pm. Yeah, is it open on weekends too? Is there anything else I can help you with today? |
|---|---|
| $u_T$ | 0.2 |
| $\gamma$ | 1.4 |
| It is open from Monday to Friday from 9:30 am to 4:00 pm. On Saturday it opens at 11:00 am to 5:30 pm and on Sunday it is open at 9:00am to 3:30. Is there anything else I can help you with? |