Reranking *Over*generated Responses for End-to-End Task-Oriented Dialogue Systems

Songbo Hu, Ivan Vulić, Fangyu Liu, Anna Korhonen
Language Technology Lab
University of Cambridge, UK
{sh2091,iv250,fl399,alk23}@cam.ac.uk

Abstract

End-to-end (E2E) task-oriented dialogue (ToD) systems are prone to falling into the so-called 'likelihood trap', resulting in generated responses that are dull, repetitive, and often inconsistent with dialogue history. Comparing ranked lists of multiple generated responses against the 'gold response' (from evaluation data) reveals a wide diversity in response quality, with many good responses placed lower in the ranked list. The main challenge, addressed in this work, is then how to reach beyond *greedily generated* system responses, that is, how to obtain and select such high-quality responses from the list of *overgenerated* responses at inference *without the availability* of the gold response. To this end, we propose a simple yet effective reranking method that aims to select high-quality items from the lists of responses initially overgenerated by the system. The idea is to use any sequence-level (similarity) scoring function to divide the semantic space of responses into high-scoring versus low-scoring partitions. At training, the high-scoring partition comprises all generated responses whose similarity to the gold response is higher than the similarity of the greedy response to the gold response. At inference, the aim is to estimate the probability that each overgenerated response belongs to the high-scoring partition, given only previous the dialogue history. We validate the robustness and versatility of our proposed method on the standard MultiWOZ dataset: it improves a state-of-the-art E2E ToD system by 2.0 BLEU, 1.6 ROUGE, and 1.3 METEOR scores, achieving new peak results. Additional experiments on the BiToD dataset and human evaluation further ascertain the generalisability and effectiveness of the proposed framework.

Keywords: task-oriented dialogue systems, natural language generation, contrastive learning

1. Introduction

Task-oriented dialogue (ToD) systems (Williams and Young, 2007; Young et al., 2013) have received increasingly intensified research interest, as they can assist humans with or automate many tasks effectively, thereby contributing to technological expansion and inclusion (Raux et al., 2003; El Asri et al., 2017; Budzianowski et al., 2018; Laranjo et al., 2018). The natural language generation (NLG) module, also dubbed *response generation*, is a critical component of any ToD system. Besides the necessary requirement to maintain semantic coherence during conversation, NLG also impacts user experience and satisfaction with a system.

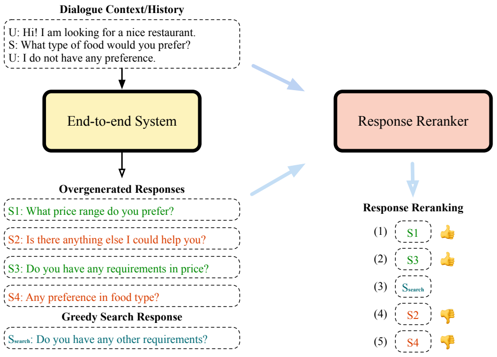
Enabled by the recent advances in pretrained language models (PLMs) (Radford et al., 2019; Raffel et al., 2020), now a *de facto* approach to NLG is fine-tuning autoregressive language models on a domain-specific dialogue dataset (Lin et al., 2020; Peng et al., 2022). However, this approach still suffers from several crucial issues. **1)** Standard autoregressive models over-rely on local context (Khandelwal et al., 2018; Sun et al., 2021), whereas many desirable properties of a dialogue response, such as consistency or coherence, can be captured only when taking into account dialogue history (Zaib et al., 2021). **2)** Autoregressive LMs make predictions conditioned on the ground truth during training but on their own predictions during decoding, creating a disparity known as 'exposure bias' (Bengio et al., 2015; Ranzato et al., 2016; Du and Ji, 2019). **3)** Finally, decoding dialogue responses from PLMs can easily fall into the so-called 'likelihood trap' (See et al., 2019a; Zhang et al., 2021); here, high-likelihood (i.e., low-perplexity) sequences produced by greedy decoding or beam search tend to be dull and repetitive (See et al., 2019b). Truncated sampling methods, such as top-k (Fan et al., 2018), nucleus (Holtzman et al., 2020), and typical sampling (Meister et al., 2022) also tend to produce text with inconsistencies, hallucinations, factual errors, or commonsense issues (Massarelli et al., 2020; Dou et al., 2022; Krishna et al., 2021; Dziri et al., 2022).

To tackle these issues, we propose a *post-generation reranking method for ToD*. The focus is on end-to-end (E2E) ToD systems, where NLG is modelled as a sequence-to-sequence problem. In particular, an E2E ToD system utilises a neural model such as T5 (Raffel et al., 2020) or BART (Lewis et al., 2020) to generate a surface form response conditioned on dialogue history and other context (e.g. dialogue domain ontology).

The method, illustrated in Figure 1, reranks a set

Fangyu Liu is now at Google Deepmind.

LREC-COLING 2024, pages 13970–13991
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



Dialogue Context/History
U: Hi! I am looking for a nice restaurant.
S: What type of food would you prefer?
U: I do not have any preference.
End-to-end System
Response Reranker
Overgenerated Responses
S1: What price range do you prefer?
S2: Is there anything else I could help you?
S3: Do you have any requirements in price?
S4: Any preference in food type?
Greedy Search Response
Ssearch: Do you have any other requirements?
Response Reranking
(1) S1
(2) S3
(3) Ssearch
(4) S2
(5) S4

Figure 1: An illustration of our proposed reranking method. S: System; U: User. A reranking model is trained to rank a set of overgenerated responses from an end-to-end ToD system solely based on dialogue context/history. According to a predefined scoring function (e.g., cosine similarity, BLEU), a good candidate should have a score higher than that of the greedy search response. After reranking, the output of the base ToD E2E system is steered towards higher-quality responses.

of responses generated by any E2E ToD system solely based on the preceding dialogue context. Inspired by prior work on conversational representation learning (Mehri et al., 2019; Humeau et al., 2020; Vulić et al., 2021b), we fine-tune any input PLM (e.g., BERT, RoBERTa) into a conversational encoder that learns fine-grained interactions between target responses and dialogue context.[1] The method is designed as a two-stage approach, see Figure 3 later. In **Stage 1** we adaptively fine-tune the input PLM according to a specified *scoring function* (e.g., cosine similarity, BLEU) and then use it to divide the semantic space (i.e., corresponding sets of generated responses) into *high-scoring* and *low-scoring* partitions based on their similarity to the gold response (according to the scoring function). Subsequently, in **Stage 2** such a specialised dialogue encoder allows reranking the generated responses based on discriminative classification or similarity-based retrieval, without leveraging the gold response. In turn, this enables us to run reranking without gold responses at inference.

In our main experiments on the standard MultiWOZ 2.0 dataset (Budzianowski et al., 2018), we run the proposed method on top of the state-of-the-art (SotA) MinTL E2E ToD system (Lin et al., 2020), and achieve consistent gains with different underlying PLMs: relying on cosine similarity as the scoring fuction, our rerankers achieve new SotA results of 20.0 (↑2.0) BLEU, 32.8 (↑1.6) ROUGE, and 36.9 (↑1.3) METEOR on MultiWOZ. Further, using the actual evaluation metric also as the scoring criterion in Stage 2, we can push performance to 20.3 (↑2.3) BLEU, 33.6 (↑2.4) ROUGE, and 40.0 (↑4.4) METEOR. Ablation studies, additional experiments on the BiToD dataset (Lin et al., 2021), and human-based evaluation further verify the usefulness of the proposed method. The code is available online at `https://github.com/cambridgeltl/response_reranking`.

2. Related Work

We focus on improving NLG performance of E2E ToD systems (Wen et al., 2017; Bordes et al., 2017; Lei et al., 2018; Lin et al., 2020), proposing a post-generation reranking method which operates on the E2E system's outputs.

Post-Generation Reranking. It has been well-studied in the Machine Translation (MT) community. Noisy Channel Modelling (Ng et al., 2019; Yee et al., 2019) is a widely used reranking scheme for NMT, parameterising the noisy channel probability with a seq2seq model. Rerankers have also been implemented with an RNN language model (Gulcehre et al., 2017), an energy-based model (Bhattacharyya et al., 2021), and masked language models (Salazar et al., 2020; Liu and Liu, 2021).

In dialogue, reranking methods were mostly investigated for open-ended systems, aiming to increase their response diversity (Sordoni et al., 2015; Li et al., 2016; Shao et al., 2017), to improve conversational 'engagingness' by integrating human feedback (Gao et al., 2020), and to enhance fluency and semantic correctness (Baheti et al., 2020). However, post-generation reranking has not been as widely explored in the ToD context. Compared to open-ended systems, generated outputs from ToD systems are highly semantically similar, adding a substantial challenge to reranking, reaching beyond simple topic shifts of the responses. Notably, rerankers based on convolution (Wen et al., 2015), RNN (Dušek and Jurčíček, 2016), RoBERTa (Harkous et al., 2020), and cross-attention (Juraska and Walker, 2021) were proposed, which all crucially assume access to the ground truth dialogue act representation, while we do not assume its availability.

Response Selection for ToD. These methods retrieve a set of response candidates and subsequently select the most likely one (according to a matching function) as a final response (Ritter et al., 2011). Different matching models have been proposed to measure the matching degree between a dialogue context and a response candidate, and rank the candidates accordingly (Wu et al., 2017; Zhou et al., 2018; Weston et al., 2018; Lu et al.,

---

[1] Unlike prior work, we do not use any task annotation directly for fine-tuning (e.g., prior work relied on intent labels to specialise PLMs towards particular intent detection tasks).

2019; Gu et al., 2019; Su et al., 2021; Henderson et al., 2020; Humeau et al., 2020, *among others*). Unlike prior work, which typically ranks a set of predefined system response candidates, our post-generation reranking method combines generation-based and retrieval-based methods. Moreover, while previous work (*e.g.* Weston et al., 2018; Dinan et al., 2019; Kim et al., 2020) augmented the dialogue context with retrieved knowledge *before* generation, our method is a post-generation reranking method. In particular, our rerankers operate on a set of over-generated (and thus semantically close) responses; we thus need to capture very subtle nuances between different response candidates.

Contrastive Learning for NLG. Contrastive learning (CL) (Chopra et al., 2005; Schroff et al., 2015; Chen et al., 2020; He et al., 2020) has been widely used in NLP for word-level (Mikolov et al., 2013; Vulić et al., 2021a; Liu et al., 2021b) and sentence representation learning (Reimers and Gurevych, 2019; Wu et al., 2020; Meng et al., 2021; Liu et al., 2021a; Gao et al., 2021). Beyond representation learning, other work applies CL to open-ended text generation (Krishna et al., 2022; Su et al., 2022). However, as posited by Krishna et al. (2022), such a method may not be directly applicable to other generation tasks with a more constrained output space (e.g. NLG for ToD). For constrained generation tasks, Liu and Liu (2021) apply CL to post-generation reranking for abstractive summarisation, and An et al. (2022) use CL for five generation tasks, but none of them relates to dialogue.

3. Post-Generation Response Reranking

Motivation: An Oracle Experiment. In the 'oracle' experiment, where we assume the availability of the ground truth response, we first focus on examining the diversity of candidate responses generated by the underlying E2E ToD model, which would outline the potential of post-generation reranking. We rerank the set of 20 'oversampled' responses from an E2E ToD system, using their sentence-level BLEU-based similarity to the ground truth. As revealed by Figure 2, when generating the 20 responses with the nucleus sampling method (Holtzman et al., 2020), we can find responses in those sets that are of much higher-quality (as measured by BLEU) as well as of much lower-quality than the standard *greedy* response. It is possible to improve BLEU up to 16.2 points if we always select the best response (according to BLEU) from the 20-response set.

The crucial issue is that at real-world 'non-oracle' inference we cannot leverage such ground truth responses. Oservations from the oracle experiment indicate that: (i) there is ample room for improvement in NLG via reranking offering empirically driven motivation for our novel reranking methods, while (ii) we need to disentangle the critical dependency on the ground truth response from the reranking process.

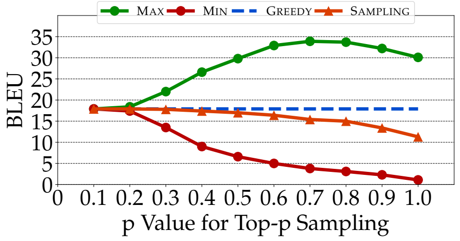


Figure 2: Corpus BLEU scores for MinTL on the MultiWOZ *dev* set in the 'oracle' experiment. Nucleus top-p sampling with different p values. Max (Min) performance is achieved by reranking 20 overgenerated samples based on their sentence BLEU score with the ground truth response and selecting the response with the maximum (or minimum) score. Greedy search is the standard decoding strategy. See also §4 for the short descriptions of different methods.

Response Reranking: Preliminaries. The task is similar to response selection (Wu et al., 2017; Henderson et al., 2019; Humeau et al., 2020, *among others*). Given a dataset with n examples $\mathcal{D} = \{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \cdots, \mathcal{D}^{(n)}\}$, each example $\mathcal{D}^{(i)} \in \mathcal{D}$ contains the pair $(c^{(i)}, r^{(i)})$, where $c^{(i)}$ is the dialogue context and $r^{(i)}$ is the response; $\mathbf{c}^{(i)}$ and $\mathbf{r}^{(i)}$ denote their respective representations/embeddings. During training, the task is to learn a scoring function $s(\cdot, \cdot)$ that assigns a matching score for any context–response pair. At inference, response reranking involves a dialogue model $P_{\mathrm{MLE}}(r \mid c)$ and an evaluation metric $\mathrm{M}(\cdot, \cdot)$. Given the context c, we sample a set of responses $R = \{r_1, r_2 \dots r_j\}$ from $P_{\mathrm{MLE}}(r \mid c)$. The task of a response reranker is to assign a score $s(\cdot, \cdot)$ for each context–response pair and select a response based on this score, e.g. $\mathrm{argmax}_{r \in R}\, s(\mathbf{c}, \mathbf{r})$. Response reranking is tasked to improve the evaluation score $\mathrm{M}(c, r)$.

3.1. Methodology

An effective response reranker for ToD should capture *subtle* differences among a set of *highly similar* candidates generated by a fine-tuned E2E ToD model. This setup is considerably more difficult than selecting the best response from randomly sampled confounders from a dialogue dataset (Henderson et al., 2019; Gunasekara et al., 2019). In a preliminary experiment, we followed

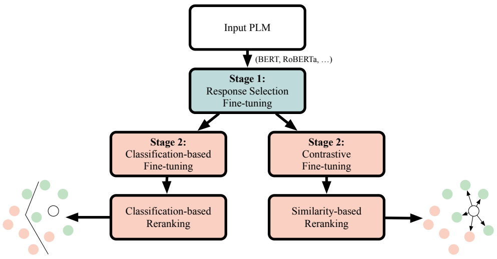


Figure 3: An overview of the two-stage reranking process. **Stage 1** fine-tunes any input PLM into a response selection model with in-domain data. **Stage 2** further fine-tunes the Stage 1 model into a response reranking model, more sensitive to fine-grained interaction between dialogue context and candidate responses.

the standard response selection setup (Wu et al., 2017; Zhou et al., 2018; Gu et al., 2019) and aimed at distinguishing between the ground truth positive example from randomly sampled negatives. We found out that those baselines perform well on the response selection task but achieve near-random performance in our response reranking task focused on highly semantically similar candidate responses. Therefore, in what follows, we propose a novel fine-tuning framework to deal with the much more challenging (re)ranking scenario.

Method in a Nutshell. We train a generative E2E dialogue model $P_{\text{MLE}}(r \mid c)$ on a dialogue dataset $\mathcal{D}$. Subsequently, for each training example (c, r) in the training set, we sample a set of responses $\mathcal{R} = \{r_1, r_2 \ldots r_j\}$ from $P_{\text{MLE}}(r \mid c)$, where j denotes the number of over-generated responses. For each $r_k \in \mathcal{R}$, we calculate its score based on a *scoring function* $s_k = s(\mathbf{r}_k, \mathbf{r})$, where $\mathbf{r}$ is the representation of the ground truth response. Unless stated otherwise, the default scoring function is defined as the cosine similarity based on a general-purpose sentence encoder.[2] We then cluster the sampled responses $\mathcal{R}$, based on their respective scores and a defined *thresholding procedure* (see §3.3), into a high-scoring set $\mathcal{R}_{high}$ and a low-scoring set $\mathcal{R}_{low}$. During training, the reranking model aims to directly capture the distinction between $\mathcal{R}_{high}$ and $\mathcal{R}_{low}$. At inference, the reranking model scores and ranks a candidate response r_k based on the *probability that the generated response is drawn from the high-scoring set*, namely $P(r_k \in \mathcal{R}_{high} \mid c)$. Again, we stress that our reranking model does not require access to the ground truth during inference. Instead, the conditional probability serves as a 'proxy' function, indicating the likelihood that the candidate is indeed a valid response. This way, we make a transition from 'oracle'-based training with ground truth to the 'non-oracle' inference.

Following Vulić et al. (2021b), we propose a two-stage fine-tuning procedure, with two types of reranking models in the second stage: a *classification-based* model and a *similarity-based* model, as illustrated in Figure 3. The framework can be applied on top of any input (Transformer-based) encoder $\mathbf{e} = enc_\theta(t)$ parameterised by θ, which encodes textual input t into a sentence embedding. Unless stated otherwise, we use BERT(-base) (Devlin et al., 2019) as our default encoder.

3.2. Stage 1: Response Selection

In Stage 1, we conduct adaptive fine-tuning in the response selection task (Vulić et al., 2021b), which transforms the input PLM into a text encoder that is better aligned with the end-task (Ruder, 2021) of response reranking. We rely on a standard cross-encoder architecture that directly models the interaction between the context and the candidate responses. Each data example is a tuple (c, r, l), where $l \in \{0, 1\}$ is a binary label indicating if r is the ground truth response to c. In fact, for each dialogue $(c^{(i)}, r^{(i)}) \in \mathcal{D}$, we construct a positive example $(c^{(i)}, r^{(i)}, 1)$. We then randomly sample a set of N_r negative responses $\mathcal{R}_{i,-}$ per each positive response $r^{(i)}$ from other tuples following prior work on response selection: for each $r^{(j)} \in \mathcal{R}_{i,-}$ it holds $i \neq j$, and we construct final negative samples as follows: $(c^{(i)}, r^{(j)}, 0)$.

The goal of Stage 1 is to fine-tune the input PLM/encoder into a statistical model parameterised by θ to compute $P_\theta(l|c, r)$. Given a training example (c, r, l), the model is trained to predict the correct label by encoding the concatenation of a context response pair $[c, r]$. To this end, the representation of the "`[CLS]`" token is subsequently projected down to two logits and passed through a softmax layer to form a Bernoulli distribution indicating the positive (1) or the negative (0) label.

3.3. Stage 2: Response Reranking

Each data entry for response reranking in Stage 2 is again a tuple (c, r, l), where $l \in \{0, 1\}$ is a binary label. We construct those data entries as follows. First, for each dialogue item $(c, r) \in \mathcal{D}$, we generate a set of responses $\mathcal{R} = \{r_1, r_2 \ldots r_j\}$ and a greedy search response r_{search}. We then calculate a pair-wise score s_k between each generated response $r_k \in \mathcal{R}$ and the ground truth response r, relying on some *scoring function* (e.g., cosine similarity between their sentence embeddings). Similarly, we calculate a score s_{search} for the greedy search response r_{search}. The score s_{search} is used as a local *threshold value* that splits the set of generated responses into positive (i.e., 'high-quality')

---

[2] We use the `all-mpnet-v2` (Reimers and Gurevych, 2019) as a robust, efficient and high-performing choice.

and negative ('low-quality') responses as follows.

If $s_k \geq s_{search}$, we add the generated response r_k to the high-quality set $\mathcal{R}_{high}$; if $s_k < s_{search}$, we add r_k to the set $\mathcal{R}_{low}$.[3] Since the cardinality of the two sets may differ, we downsample the larger set to the size of the smaller one: $min(|\mathcal{R}_{high}|, |\mathcal{R}_{low}|)$. Following that, for each $r_k \in \mathcal{R}_{high}$, we construct a positive example for fine-tuning $(c, r_k, 1)$; for each $r_k \in \mathcal{R}_{low}$, we construct a negative example $(c, r_k, 0)$. We construct such examples from the entire training set, and they are then used for two types of reranking: *classification-based* and *similarity-based*, described in what follows, with additional illustrations in Figure 6 in Appendix A.

Classification-Based Reranking. This procedure is identical to Stage 1. However, the reranking models now learn to rerank overgenerated (and semantically similar) responses according to positive and negative examples corresponding to respective sets $\mathcal{R}_{high}$ and $\mathcal{R}_{low}$. Given a training example (c, r, l), the encoder first encodes the $[c, r]$ and the contextualised representation of the "`[CLS]`" token is subsequently used to compute $P_\theta(l|c, r)$ as in Stage 1 using the standard cross-encoder architecture (*e.g.* Wolf et al., 2019; Urbanek et al., 2019). At inference, given $\mathbf{c}$ and a set of generated responses $\mathcal{R}$, we rank and select the final response based on its score: $\text{argmax}_{\mathbf{r} \in \mathcal{R}} P(l = 1|c, r)$.

Similarity-Based Reranking. Similarity-based classification has demonstrated promising results in intent detection for ToD (Zhang et al., 2020a; Vulić et al., 2021b) and other NLP tasks (Sarwar et al., 2022; Kassner and Schütze, 2020), particularly when data is scarce. We thus also propose a similarity-based reranker in Stage 2, based on contrastive fine-tuning and KNN retrieval.

The aim is to fine-tune the input encoder so that it encodes all context-response pairs from $\mathcal{R}_{high}$ into coherent clusters, clearly separated from low-scoring pairs from $\mathcal{R}_{low}$. Here, we utilise the label l during training only implicitly, allowing us to formulate reranking as a sentence similarity task. In particular, for a training example (c, r, l), the encoder first encodes $[c, r]$, where the encoding $\mathbf{e} = enc_\theta([c, r])$ is created via mean-pooling over the constituent subwords' embeddings.

We use the standard Triplet Loss (Schroff et al., 2015). For any pair of examples within a batch $(enc_\theta([c^{(i)}, r^{(i)}]), l^{(i)})$ and $(enc_\theta([c^{(j)}, r^{(j)}]), l^{(j)})$, the encoder parameters θ are optimised (i) to reduce the cosine distance between encodings of the pairs with the same label $l^{(i)} = l^{(j)}$, and (ii) to increase the distance otherwise. After Stage 2, response scoring in the specialised encoder space enc_{S_2} is then performed via similarity-based KNN inference (Zhang et al., 2020a; Vulić et al., 2021b), using a subset of training examples as *anchors*. For all anchors $(c^{(i)}, r^{(i)}, l^{(i)})$, we compute their encodings $\mathbf{e}^{(i)} = enc_{S_2}([c^{(i)}, r^{(i)}])$ in advance. For any candidate-response pair, we obtain its encoding $\mathbf{e} = enc_{S_2}([c, r])$. We retrieve a set of k nearest anchors from the full set of anchors. The scoring function $s(\cdot, \cdot)$ is defined as the proportion within the k nearest anchors with a positive label. We select the final response as follows: $\text{argmax}_{r \in \mathcal{R}} s(c, r)$.

4. Experimental Setup

Our main experiments focus on the standard multi-domain MultiWOZ ToD dataset (Budzianowski et al., 2018): in particular on its 2.0 version.

Baseline E2E System. The underlying E2E ToD system is MinTL (Lin et al., 2020), as a publicly available SotA model. It jointly learns dialogue state tracking and response generation with pretrained seq2seq models.[4] However, we note that the proposed reranking method can be applied to any E2E dialogue system with autoregressive response generation (*e.g.,* Wen et al., 2017; He et al., 2022).

Evaluation Metrics. Following the standard MultiWOZ setup, we use the corpus BLEU score (Papineni et al., 2002) as our primary evaluation metric, and all the scores are computed with delexicalised utterances based on the DAMD system (Zhang et al., 2020b).[5] We also report ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) as two other standard NLG evaluation metrics.

Input PLMs for Reranking. Our method can be implemented with any Transformer-based (Vaswani et al., 2017) PLM. To analyse the impact of the input PLM (see Figure 3), we experiment with several popular PLMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and their distilled versions (Sanh et al., 2019). We additionally experiment with supervised sentence encoders: SimCSE (Gao et al., 2021) and other popular

---

[3] There are other options to split the candidate responses into the sets $\mathcal{R}_{high}$ and $\mathcal{R}_{low}$ (e.g., selecting the top N% responses for $\mathcal{R}_{high}$). However, our proposed method is hyperparameter-free and is also conditioned on the score of the greedy response; greedy search is the default decoding strategy of many standard E2E ToD systems (Lin et al., 2020).

[4] MinTL (Lin et al., 2020) was the SotA system for E2E NLG on the MultiWOZ leaderboard until the recently published GALAXY system (He et al., 2022) surpassed its performance by a 0.2 BLEU score. See the MultiWOZ leaderboard at `https://github.com/budzianowski/multiwoz`.

[5] In the case of delexicalised dialogues, all the slot values in the context and responses are replaced a predefined placeholder (e.g. *[value_name] is an [value_price] [value_food] restaurant on the [value_area] . do you need to know more ?*).

encoders from the sentence-transformers (i.e., SBERT) repository (Reimers and Gurevych, 2019). Table 8 in the appendix lists all the input models we use along with their checkpoints from the HuggingFace repository (Wolf et al., 2020).

Hyperparameters and Optimisation. The default decoding strategy for MinTL is the greedy search. In our reranking experiments, unless stated otherwise, we over-generate 20 responses with nucleus sampling (Holtzman et al., 2020) from the top-0.7 portion of the probability mass, a standard choice.

We implement all reranking models via the SBERT repository (Reimers and Gurevych, 2019), which is built on top of the HuggingFace repository (Wolf et al., 2020). Table 6 in the appendix lists the search set of hyperparameters (which differ from the default SBERT-suggested values), along with the finally set values. The grid search was conducted on the dev set, based on BLEU.[6]

Model Variants and Baselines. We experiment with several model variants enabled by the proposed two-stage pipeline (see Figure 3):

PLM+S1+S2. This variant refers to the full pipeline, where PLM is any input PLM from Table 8. Stage 1 (S1) fine-tuning can be based on either lexicalised dialogues (*S1:lex*) or delexicalised (*S1:delex*) dialogues. After S1, we can further fine-tune the 'S1' encoders via the classification-based or the similarity-based approach: *S2:class* and *S2:sim*. For instance, the configuration *BERT+S1:delex+S2:sim* denotes the use of BERT as the input PLM, with delexicalised dialogues in Stage 1, and similarity-based Stage 2.

PLM+S2. This group is fine-tuned only relying on Stage 2 approaches, skipping Stage 1.

PLM+S1. This group is fine-tuned only for response selection with in-domain data, ignoring S2.

PLM. This variant refers to using out-of-the-box sentence encoders in the response reranking task. Since classification-based reranking requires a fine-tuned task-specific classification head, we only run our experiments with similarity-based reranking.

We also compare against two standard decoding strategies. **1) Greedy**. Greedy search has been widely used as the default decoding strategy NLG, also by the base MinTL system (Lin et al., 2020).[7] **2) Sampling.** As mentioned, we apply nucleus sampling (Holtzman et al., 2020) to sample responses from the top-0.7 portion of the probability mass.

Variant	Selection	Reranking		
	R@1	B	R	M
Random Sampling	5.0	15.8	27.3	31.0
Greedy	–	18.0	31.2	35.6
BERT	–	17.0	29.4	33.6
SimCSE	–	16.7	29.0	33.2
all-mpnet	–	16.0	27.6	31.8
BERT+S1:delex	51.0	16.7	39.3	33.8
BERT+S1:lex	77.2	17.1	29.7	34.3
DRoB+S1:delex	48.0	16.6	29.0	33.4
DRoB+S1:lex	74.4	16.6	29.6	34.5

Table 1: Performance of representative out-of-the-box sentence encoders and response selection models on the standard Response **Selection** task (R@1 = Recall@1), and on the final Response **Reranking** task, relying on the MultiWOZ test set. Similarity-based reranking without S2 fine-tuning is reported. B=BLEU; R=ROUGE; M=METEOR; DRoB=DistilRoBERTa. Additional results (incl., classification-based reranking) with more input models are available in Table 10 in Appendix C.

5. Results and Discussion

Before delving into the main results, we investigated the capability of standard response selection techniques from the literature to select the best response from the 'overgenerated set', with the scores summarised in Table 1: they reveal that the standard approaches are inadequate for our task, all scoring below the *Greedy* baseline.

5.1. Main Results

The main results are summarised in Table 2. They suggest that our classification-based reranker yields 20.0 BLEU, outperforming the stronger *Greedy* baseline by 2.0 points. Similar gains are achieved by our similarity-based variant. The comparison of results in Tables 2 and 1 further indicates the inadequacy of standard response selection methods in the reranking task, and the importance of Stage 2 fine-tining: our two-stage reranking framework provides consistent and robust gains over the baselines across the board. Delving deeper into the model performance through ablation experiments, reported in Table 3, isolates the critical components responsible for the strong performance.

---

[6] 128 is the maximum batch size with BERT base and RoBERTa for Stage 2 fine-tuning. Following Reimers and Gurevych (2019); Vulić et al. (2021b), we use the AdamW optimiser (Loshchilov and Hutter, 2019) in the default SBERT setting: the learning rate is 2e-5; warmup of 0.1 and linear decay; the weight decay rate is 0.01. We rely on the triplet loss variant of Hermans et al. (2017): this is `BatchAllTripletLoss` in the SBERT repo; see the documentation and the paper for further details.

[7] Greedy search and beam search are used as the default decoding methods for many SotA E2E systems (Lin et al., 2020, 2021; He et al., 2022) as they typically outperform sampling algorithms in terms of BLEU. This finding has also been corroborated by our 'oracle' experiment; see Figure 2.

Variant	**BLEU**	**ROUGE**	**METEOR**
Baselines			
Sampling	15.8	27.3	31.0
Greedy	18.0	31.2	35.6
BERT **Classification-based**			
+S2	19.4	32.1	36.4
+S1:delex+S2	19.3	32.3	36.3
+S1:lex+S2	19.3	32.1	36.2
quora-distilroberta **Classification-based**			
+S2	19.6	32.0	36.1
+S1:delex+S2	**20.0**	**32.8**	**36.9**
+S1:lex+S2	19.8	32.6	36.7
BERT **Similarity-based**			
+S2	18.6	30.8	34.8
+S1:delex+S2	**19.6**	**32.0**	**36.5**
+S1:lex+S2	19.1	31.7	36.0

Table 2: Reranking performance with selected model variants based on 20 over-generated responses from MinTL. Full results with other PLMs and variants are available in Table 11 in Appendix C.

Variant	**BLEU**
Classification-based	
quora-distilroberta+S1:delex+S2	20.0
- self-generated positives	13.7 (↓6.3)
- multiple positives	19.1 (↓0.9)
- cross-encoders (+ bi-encoders)	15.4 (↓4.6)
Similarity-based	
BERT+S1:delex+S2	19.6
- self-generated positives	15.2 (↓4.4)
- multiple positives	19.2 (↓0.4)

Table 3: Ablations on the two best-performing reranking models. We can replace cross-encoders in the classification-based model with the bi-encoder architecture: it encodes contexts and responses separately sharing the encoder's weights; it was trained via the Softmax loss (Reimers and Gurevych, 2019).

Self-Generated Positives: Previous work (Krishna et al., 2022) utilises self-generated sentences only to construct negative examples for contrastive learning. Put simply, the model in prior work is trained to select the provided ground truth among self-generated examples. However, for our response reranking task, we need to select the best response from a set where all the items are self-generated. The results suggest that, modelling self-generated responses as positives in S2 (i.e., creating the set $\mathcal{R}_{high}$) is crucial for the reranking effectiveness. The performance degrades considerably for both S2 variants without self-generated positives in S2. The results further suggest that incorporating multiple positive pairs into the same batch yields slight performance gains for both S2 variants.

S2 Scoring↓ / Evaluation→	BLEU	ROUGE	METEOR
Greedy	18.0	31.2	35.6
Similarity	19.3	32.3	36.3
BLEU	20.3	33.2	37.2
ROUGE	**20.7**	**33.6**	37.6
METEOR	18.2	33.4	**40.0**

Table 4: Reranking performance of *BERT+S1:delex+S2* with Classification-based S2; the scores with Similarity-based S2 are similar, see Table 12 in Appendix C. For the models with ROUGE and METEOR as scoring functions, we perform model selection based on the best ROUGE and METEOR performance, respectively.

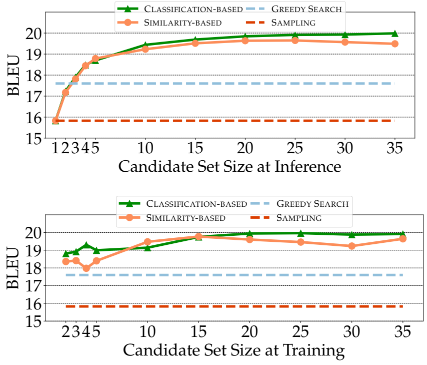


Figure 4: Impact of the candidate set size for reranking during **(Top)** inference and **(Bottom)** training on the final BLEU performance. The plots focus on the best-performing model variants from Table 2.

Cross-Encoders: Encoding context-response pairs with cross-encoders instead of using bi-encoders (Humeau et al., 2020; Henderson et al., 2020) leads to better performance. Cross-encoders are able to capture finer-grained interactions between the context and the response (Geigle et al., 2022), which is pivotal for the response reranking task dealing with subtle variations in the semantically close candidate responses.[8] Further, cross-encoders also enable our similarity-based reranking models.

5.2. Further Analysis

We now analyse other important aspects of the proposed reranking framework, running a series of side experiments, with additional analyses of (arguably) lower importance available in Appendix C.

Impact of the Input Encoder. Figure 5 shows the reranking performance with different encoders. Interestingly, the distilled PLMs achieve performance

---

[8] Cross-encoders usually perform better than bi-encoders with the caveat of reduced efficiency (Urbanek et al., 2019), but they are typically used exactly in reranking contexts (Geigle et al., 2022; Li et al., 2022) similar to ours.

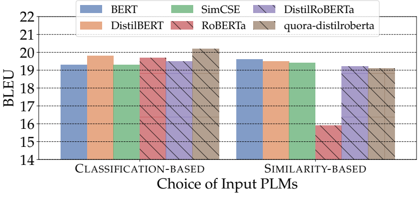


Figure 5: Reranking performance for different input PLMs trained with the entire fine-tuning pipeline (PLM+S1:delex+S2). Full results are in Appendix C.

Variant	B	R	M
Beam Search	45.8	45.5	50.7
Sampling	43.0	42.9	48.4
BERT+S1:delex+S2:class	46.3	45.9	51.4
BERT+S1:delex+S2:sim	45.9	45.4	51.3

Table 5: Reranking performance based on 20 overgenerated responses on English BiToD. B=BLEU; R=ROUGE; M=METEOR.

which is on-par with larger models; see also Figures 8-9 in Appendix C. Further, sentence encoders such as SimCSE and quora-distilroberta do not yield any gains over the other encoders.

Evaluation Metrics as Scoring Functions. Table 4 and Table 12 in Appendix C indicate that the gains are consistent across all three automatic evaluation metrics. Moreover, using any of the three metrics as the scoring function s in S2–dividing generated samples into the sets $\mathcal{R}_{high}$ and $\mathcal{R}_{low}$ (see §3.3)—yields gains on all the other metrics as well. Naturally, higher scores per each evaluation metric are typically achieved when the same metric is used as the scoring function in S2. However, such a setup might provide artificially inflated metric-specific performance; a better indicator of effectiveness and robustness is the evaluation metric-agnostic model variant used throughout the paper (i.e., the row *Similarity* in Table 4), which relies on the standard cosine similarity to do partitioning in Stage 2.

Impact of the Candidate Set Size. Figure 4 plots the reranking performance conditioned on a varying number of overgenerated responses during inference.[9] The curves indicate that **1)** both reranking variants already outperform the Greedy search baseline and achieve SotA performance when the candidate set spans only three candidates; **2)** after the sharp increase in performance for the sizes 1-10, further increase in the candidate set size offers diminishing returns as performance saturates. In addition, Figure 4 (bottom) demonstrates that the reranking models outperform the Greedy baseline with only two overgenerated responses for training.

Another ToD Dataset. To test the generalisability of the proposed method, we also run experiments on the English portion of the BiToD dataset (Lin et al., 2021). The experimental setup is described in Appendix D, while the results are summarised in Table 5. Our reranking framework again yields gains over the baselines, verifying its robustness, but the gains are now less pronounced. Delving deeper into the roots of this result, we attribute this to BiToD's data properties combined with the baseline system: mT5 (Xue et al., 2021), resulting in the lack of syntactic and semantic variability in MinTL's generated outputs on BiToD. On average, there are only 10.4 unique utterances within the set of 20 overgenerated items, compared to 17.8 for MultiWOZ, with cases where all the 20 generated responses are identical, which leaves meagre or no room for further improvement via reranking.

Human Evaluation. User satisfaction is always the ultimate goal of developing ToD systems (Ji et al., 2022). We thus additionally evaluate with human subjects, with the details on the setup in Appendix E. We follow suggestions from prior work (Fomicheva et al., 2021), and conduct comparative 'A/B' tests with 6 subjects, each scoring 100 dialogues. Each test item contains a dialogue context and three randomly ordered outputs from greedy search, classification-based reranker, and similarity-based reranker; the human participant's task is to indicate pairwise preferences among the three outputs. The results indicate: (i) a 49.5% (0.5% less) preference for classification-based reranker over the greedy search (0.28 Fleiss' Kappa); (ii) a 57.8% preference for similarity-based reranker over the greedy search baseline (0.26); (iii) a 55.8% preference for similarity-based reranker over the classification-based reranker (0.16). Overall, we see a slight preference towards similarity-based rerankers.[10] However, we did not observe a strong inter-annotator agreement, as measured by Fleiss' Kappa (Fleiss, 1971), as the scoring task is considered highly subjective, and most responses are similar and difficult to distinguish, even for humans (e.g., see an example in Figure 10). This evaluation difficulty again reflects the difficulty of our proposed reranking task as a particularly challenging modelling scenario for neural models.

---

[9] The number of overgenerated samples for inference is critical for real-world applications: inference time scales linearly with this number; see more in the Limitations section.

[10] Following Welleck et al. (2020), we run 2-sided binomial tests indicating the significance of the human preference. See detailed results in Table 14 in Appendix E.

6. Conclusion

We proposed a novel post-generation reranking method applicable to any end-to-end (E2E) task-oriented dialogue (ToD) system. The reranking is formulated as a two-stage conversational fine-tuning procedure that transforms any input pretrained LM into a specialised in-domain reranker which can operate on the sets of generated responses from the E2E ToD system. Combined with a strong E2E ToD system, our reranking models improved E2E dialogue generation performance on standard ToD benchmarks, and achieved new state-of-the-art results on the MultiWOZ benchmark, complemented with favourable human evaluation. Our method operates at inference time, showing adaptability to the rapidly evolving paradigms of ToD systems. A promising avenue for future research is to explore its integration with large language models and in-context learning within ToD systems.

7. Limitations

One limitation of the proposed reranking is of practical nature and concerns its dependence on two expensive operations: overgeneration and reranking. In theory, the time complexity of overgeneration scales linearly with the number of outputs, similarly to the beam size for beam search. In practice, we observe that, with the HuggingFace implementation (Wolf et al., 2020), sampling 10 responses doubles the time consumption compared to sampling a single response. In addition, unlike beam search, this over-sampling can easily be parallelised for real-world applications. Our method improves the baseline even if we only generate three responses during inference (see Figure 4). Reranking is less time-demanding, and it takes ∼2 minutes on a single GPU (see Appendix B) for the full MultiWOZ test set with 20 candidates. In future work, we will explore more parameter-efficient methods for over-generation.

Our proposed reranking method is versatile and opens up many further extensions and experimentation beyond the scope and confines of this paper. For instance, we might incorporate the ordering of the self-generated responses and replace the current contrastive loss functions with other recent effective contrastive losses (Zhou et al., 2020; Liu and Liu, 2021). Furthermore, this paper only explores out-of-box dialogue generation models without further fine-tuning. However, as the comparison of absolute gains on MultiWOZ versus BiToD indicates, increasing *response diversity* leads to a better reranking model and better performance. In future work, we will put more effort on diversifying the set of overgenerated responses in order to harvest more benefits of reranking.

The current work is also limited only to experiments with the English language, also due to the lack of suitable ToD training data for other languages (Razumovskaia et al., 2022). The recent release of the Multi3WOZ dataset (Hu et al., 2023) expands the linguistic scope, additionally support model training for Arabic, French, and Turkish. With this new resource, we also plan to extend our model to languages beyond English, as well as other dialogue-generation tasks (e.g. open-domain dialogue generation).

Finally, our work again outlines the complexity and limitations of current evaluation protocols for E2E ToD and ToD in general, as well as the importance of reporting multiple automatic and human-based evaluation metrics.

8. Ethics Statement

The experimental study obtained full Ethics Approval from the University of Cambridge in advance. Our participants were recruited within the university who volunteered to join our experiment. They are students and academic staff who are proficient in English. The consent is obtained by signing a consent form. In addition, our models leverage two data sources: the MultiWOZ dataset and the pre-training data of each PLM employed in this study. Particularly, this dataset consists solely of hypothetical dialogues in which the domains and content have been restricted and predefined, minimising the risk of personal data being present. On the other hand, it is important to acknowledge that although these PLMs are publicly available, there exists a potential risk of privacy violations (Brown et al., 2022; Carlini et al., 2021).

9. Acknowledgements

This work has been supported by a Huawei research donation to the Language Technology Lab. The work of IV has been supported by a personal Royal Society University Research Fellowship (no 221137; 2022-2027). SH has been supported by Cambridge International Scholarship. FL has been supported by Grace & Thomas C. H. Chan Cambridge International Scholarship. We thank our human participants for helping with evaluation.

10. Bibliographical References

Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. 2022. CoNT: Contrastive neural text generation. In *Advances in Neural Information Processing Systems*.

Ashutosh Baheti, Alan Ritter, and Kevin Small. 2020. Fluent response generation for conversational question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 191–207, Online. Association for Computational Linguistics.

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.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1171–1179, Cambridge, MA, USA. MIT Press.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. 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 4528–4537, Online. Association for Computational Linguistics.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, page 539–546, USA. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Wanyu Du and Yangfeng Ji. 2019. An empirical comparison on imitation learning and reinforcement learning for paraphrase generation. 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 6012–6018, Hong Kong, China. Association for Computational Linguistics.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. 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.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of SIGDIAL 2017*, pages 207–219.

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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. 2022. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *Transactions of the Association for Computational Linguistics*, 10:503–521.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2321–2324, New York, NY, USA. Association for Computing Machinery.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Comput. Speech Lang.*, 45(C):137–148.

Chulaka Gunasekara, Jonathan K. Kummerfeld, Lazaros Polymenakos, and Walter Lasecki. 2019. DSTC7 task 1: Noetic end-to-end response selection. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 60–67, Florence, Italy. Association for Computational Linguistics.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam

Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy. Association for Computational Linguistics.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

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*.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi 3 WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

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*.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Juraj Juraska and Marilyn Walker. 2021. Attention is indeed all you need: Semantically attention-guided decoding for data-to-text NLG. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 416–431, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. BERT-kNN: Adding a kNN search component to pretrained language models for better QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3424–3430, Online. Association for Computational Linguistics.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 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.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. Rankgen: Improving text generation with large ranking models. *arXiv preprint arXiv:2205.09726*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y.S. Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Melbourne, Australia. Association for Computational Linguistics.

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.

Yaoyiran Li, Fangyu Liu, Ivan Vulić, and Anna Korhonen. 2022. Improving bilingual lexicon induction with cross-encoder reranking. *CoRR*, abs/2210.16953.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale N Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021a. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021b. MirrorWiC: On eliciting word-in-context representations from pretrained language models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 562–574, Online. Association for Computational Linguistics.

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.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Constructing interpretive spatio-temporal features for multi-turn responses selection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Florence, Italy. Association for Computational Linguistics.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Locally typical sampling. *arXiv preprint arXiv:2202.00666*.

Yu Meng, Chenyan Xiong, Payal Bajaj, saurabh tiwary, Paul Bennett, Jiawei Han, and XIA SONG. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. In *Advances in Neural Information Processing Systems*, volume 34, pages 23102–23114. Curran Associates, Inc.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence,* *Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 314–319. 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.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Antoine Raux, Brian Langner, Alan W. Black, and Maxine Eskénazi. 2003. LET's GO: Improving spoken dialog systems for the elderly and non-natives. In *Proceedings of EUROSPEECH 2003*.

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2022. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74:1351–1402.

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.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sebastian Ruder. 2021. Recent advances in language model fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sheikh Muhammad Sarwar, Dimitrina Zlatkova, Momchil Hardalov, Yoan Dinkov, Isabelle Augenstein, and Preslav Nakov. 2022. A neighborhood framework for resource-lean content flagging. *Transactions of the Association for Computational Linguistics*, 10:484–502.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019a. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019b. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219, Copenhagen, Denmark. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. 2021. Dialogue response selection with hierarchical curriculum learning. 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 1740–1751, Online. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. 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 673–683, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021a. LexFit: Lexical fine-tuning of pretrained language models. 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 5269–5283, Online. Association for Computational Linguistics.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021b. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1168, Online and Punta Cana, Dominican Republic. 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 Gašić, Dongho Kim, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–284, Prague, Czech Republic. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

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

Jason D. Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

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, Canwen 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.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 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.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Munazza Zaib, Quan Z. Sheng, and Wei Emma Zhang. 2021. A short survey of pre-trained language models for conversational AI - A new age in NLP. *CoRR*, abs/2104.10810.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

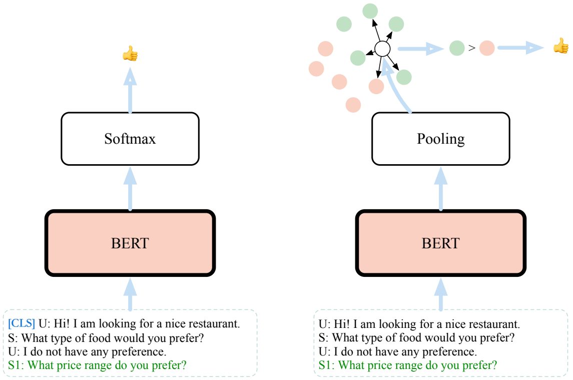
Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020a. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082, Online. Association for Computational Linguistics.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020b. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9604–9611. AAAI Press.

Mo Zhou, Zhenxing Niu, Le Wang, Zhanning Gao, Qilin Zhang, and Gang Hua. 2020. Ladder loss for coherent visual-semantic embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13050–13057.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

11. Language Resource References


Softmax
BERT
[CLS] U: Hi! I am looking for a nice restaurant.
S: What type of food would you prefer?
U: I do not have any preference.
S1: What price range do you prefer?
Pooling
BERT
>
U: Hi! I am looking for a nice restaurant.
S: What type of food would you prefer?
U: I do not have any preference.
S1: What price range do you prefer?

(a) Classification-based Reranking **(b) Similarity-based Reranking**

Figure 6: Different reranking variants in Stage 2.

A. Reranking Variants in Stage 2

Figure 6 illustrates the two proposed approaches for Stage 2 fine-tuning, with detailed descriptions available in the main paper (see §3.1).

B. Experimental Details

We run all our experiments on a single RTX 24 GiB GPU.

Table 6 lists the search set of our model hyperparameters. Unless mentioned otherwise, all the hyperparameters are set to the default values provided in the SBERT repository. For the classification-based response reranking training, the batch size is 64. For similarity-based reranking training, the batch size is 128. Both batch size values are determined as the maximum values based on our hardware (see above).

Table 8 lists all the PLMs we used in this work, along with their respective checkpoints in the Huggingface repository.

Table 9 shows time consumption of our proposed ranking models for fine-tuning and inference. The time consumption is measured based on five independent runs for the BERT-based models on the MultiWOZ dataset.

Impact of Random Initialisation. For our best-performing models, we ran five independent runs with different random seeds. The main finding is that the scores exhibit small-to-negligible variance across different runs. Namely, our best-performing classification-based model achieves the BLEU score of 19.96 ± 0.16, and our similarity-based model achieves the BLEU score of 19.59 ± 0.09.

Hyper-parameter	**Value**
Stage 1: Response Selection	
batch size	64
context window	{1, 2, 3, 4, 5}
max sequence length	128
training epoch	{1, 2, 3, 10}
candidate size	20*
Stage 2: Response Reranking	
batch size	{64, 128}
context window	{1, 2, 3, 4, 5}
max sequence length	128
training epoch	5
BatchAllTripletLoss margin	5

Table 6: Model hyper-parameters. (*)For each dialogue $(c^{(i)}, r^{(i)}) \in \mathcal{D}$, the candidate size for response selection training is 20. In other words, there are 1 positive response $r^{(i)}$ and 19 negative responses $\mathcal{R}_{i,-}$. For each $r^{(j)} \in \mathcal{R}_{i,-}$ it holds $i \neq j$.

Hyper-parameter	Value
Stage 1: Response Selection	
batch size	64
context window	3
max sequence length	128
training epoch	10
candidate size	20
Stage 2: Response Reranking	
batch size	{64, 128}
context window	3
max sequence length	128
training epoch	5
BatchAllTripletLoss mergin	5

Table 7: BiToD experiments: model hyperparameters.

Model	**HuggingFace Checkpoint**
BERT	bert-base-uncased
RoBERTa	roberta-base
DistilBERT	distilbert-base-uncased
DistilRoBERTa	distilroberta-base
SimCSE	princeton-nlp/sup-simcse-bert-base-uncased
MiniLM	sentence-transformers/all-MiniLM-L12-v2
all-mpnet-v2	sentence-transformers/all-mpnet-base-v2
quora-distilroberta	cross-encoder/quora-distilroberta-base

Table 8: Input PLMs.

C. Additional Results on MultiWOZ

To solidify our findings in this paper, we list additional experimental results which offer further empirical support for our main claims:

Table 10 shows the results of standard response selection techniques from the literature (i.e., effectively running only S1 in our pipeline) to select the best response from the sets of overgenerated candidates. All the model variants score lower than

Setup	**Time**
Stage 1: Response Selection	
Training per epoch	29:18
Inference *(full test)*	1:58
Stage 2: Response Reranking	
Training per epoch	27:10
Inference with 5,000 anchors *(full test)*	2:20
Inference with 10,000 anchors *(full test)*	4:52

Table 9: Time consumption of our proposed methods. It was computed as an average of 5 runs on a machine with a 16-core vCPU and a single RTX 24 GiB GPU.

the standard *Greedy* search baseline in the reranking task despite the fact that S1 in-domain fine-tuning increases their response selection capabilities (e.g., compare their R@1 scores versus the random baseline). Moreover, the results indicate that selecting an utterance based on delexicalised dialogue contexts is harder than with lexicalised dialogues.

Figure 8 and **Figure 9** demonstrate the reranking performance with different input PLMs, measured by the ROUGE and METEOR score, respectively (see also §5.2).

Table 11 provides the results with different input PLMs in our comparison with the full fine-tuning pipeline. From this table, sentence encoders do not provide advantages over PLMs. Table 11 can been seen as an expanded version of Figure 5, Figure 8, and Figure 9.

Table 12 displays reranking performance of the *BERT+S1:delex+S2* model variant with different scoring functions in Stage 2 for partitioning responses into the sets $\mathcal{R}_{high}$ and $\mathcal{R}_{low}$. Due to a high similarity between patterns observed with Classification-based and Similarity-based S2 reranking, we show only a Classification-based partition of the full table in the main paper: Table 4.

Table 13 provides the results with the *BERT+S1:delex+S2* model variant with varying dialogue history/context size. Both classification-based and similarity-based rerankers utilise the historical dialogue context and require at least 2 preceding historical utterances to be effective. However, there is no discernible correlation between the reranking performance and the dialogue context size further beyond. In other words, by increasing or decreasing the number of the input historical utterances, the reranking performance does not catastrophically degrade, when more than two historical utterances are available.

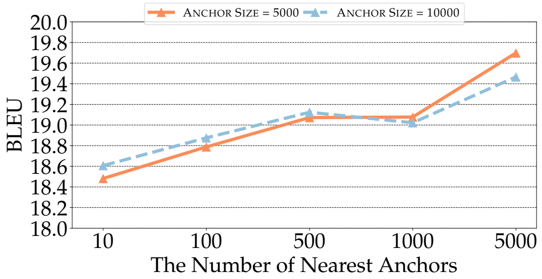


Figure 7: Performance with *BERT+S1:delex+S2:sim* using different values for the total number of anchors and the number of nearest anchors k.

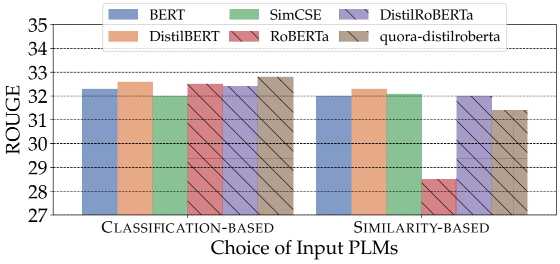


Figure 8: Reranking performance (ROUGE scores) for different input PLMs trained with the entire fine-tuning pipeline (PLM+S1:delex+S2)..

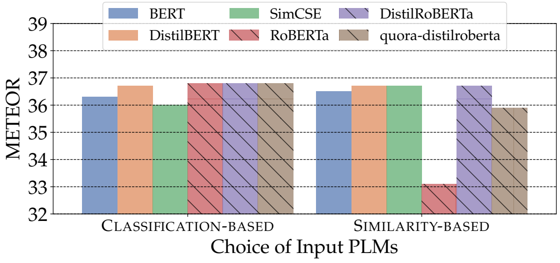


Figure 9: Reranking performance (METEOR scores) for different input PLMs trained with the entire fine-tuning pipeline (PLM+S1:delex+S2).

Similarity-Based Stage 2. There are two key hyper-parameters: the total number of anchors, and the number of nearest anchors k used to score each test example. **Figure 7** plots their impact on the final BLEU scores, indicating that the approach is fairly robust to different tested values. Note that the time consumption scales linearly with the anchor pool size (see also Table 9 in Appendix B).

D. Experimental Setup on BiToD

BiToD (Lin et al., 2021) is a bilingual (English and Chinese) multi-domain dataset for end-to-end task-oriented dialogue modelling. For our experiments,

Variant	Response Selection	Classification-based Reranking			Similarity-based Reranking		
	R@1	BLEU	ROUGE	METEOR	BLEU	ROUGE	METEOR
Random Baseline							
Random Sampling	5.0	15.8	27.3	31.0	15.8	27.3	31.0
Sentence Encoders							
BERT	n/a	n/a	n/a	n/a	17.0	29.4	33.6
SimCSE	n/a	n/a	n/a	n/a	16.7	29.0	33.2
quora-distilroberta	n/a	n/a	n/a	n/a	15.9	27.7	32.0
all-mpnet	n/a	n/a	n/a	n/a	16.0	27.6	31.8
Response Selection Models							
BERT+S1:delex	51.0	16.2	29.0	34.2	16.7	39.3	33.8
DistilRoBERTa+S1:delex	48.0	16.1	38.9	33.7	16.6	29.0	33.4
quora-distilroberta+S1:delex	50.5	16.3	28.8	33.6	16.6	29.1	33.9
BERT+S1:lex	77.2	14.5	28.8	33.2	17.1	29.7	34.3
DistilRoBERTa+S1:lex	74.4	15.0	27.7	32.4	16.6	29.6	34.5
quora-distilroberta+S1:lex	74.9	15.2	27.9	32.8	16.2	28.6	33.2

Table 10: Response selection and response reranking performance on the MultiWOZ test set with standard response selection models and out-of-box sentence encoders. n/a = non-applicable.

Variant	Classification-based Reranking			Similarity-based Reranking		
	BLEU	ROUGE	METEOR	BLEU	ROUGE	METEOR
PLMs						
BERT+S1:delex+S2	19.3	32.3	36.3	19.6	32.0	36.5
DistilBERT+S1:delex+S2	19.8	32.6	36.7	19.5	32.3	36.7
RoBERTa+S1:delex+S2	19.7	32.5	36.8	15.9	28.5	33.1
DistilRoBERTa+S1:delex+S2	19.5	32.4	36.7	19.2	32.0	36.8
BERT+S1:lex+S2	19.3	32.1 ↓	36.2 ↓	19.1 ↓	31.7 ↓	36.0 ↓
DistilBERT+S1:lex+S2	19.7 ↓	32.4 ↓	36.5 ↓	19.8	32.1 ↓	36.5 ↓
RoBERTa+S1:lex+S2	19.9	32.7	36.7 ↓	19.0	31.7	36.3
DistilRoBERTa+S1:lex+S2	19.4 ↓	32.1 ↓	36.3 ↓	18.6 ↓	31.8 ↓	37.1
Sentence Encoders						
MiniLM+S1:delex+S2	19.9	32.4	36.5	19.0	32.0	37.1
all-mpnet+S1:delex+S2	19.7	32.4	36.5	18.9	32.1	36.9
SimCSE+S1:delex+S2	19.3	32.0	36.0	19.4	32.1	36.7
quora-distilroberta+S1:delex+S2	20.0	32.8	36.9	19.1	31.4	35.9
MiniLM+S1:lex+S2	19.3 ↓	31.9 ↓	35.9 ↓	18.9 ↓	31.4 ↓	35.8 ↓
all-mpnet+S1:lex+S2	19.6 ↓	32.3 ↓	36.5	18.8 ↓	31.5 ↓	35.8 ↓
SimCSE+S1:lex+S2	19.5	32.4	36.5	19.0 ↓	31.4 ↓	35.7 ↓
quora-distilroberta+S1:lex+S2	19.8 ↓	32.6 ↓	36.7 ↓	18.3 ↓	31.3 ↓	35.9

Table 11: Response reranking models trained with the full fine-tuning pipeline with different input PLMs. ↓ denotes a lower performance compared to the counterpart model trained with delexicalised dialogues.

S2 Scoring↓ / Evaluation→	**Classification-based Reranking**			**Similarity-based Reranking**		
	BLEU	ROUGE	METEOR	BLEU	ROUGE	METEOR
Greedy	18.0	31.2	35.6	18.0	31.2	35.6
Similarity	19.3	32.3	36.3	19.6	32.0	36.5
BLEU	20.3	33.2	37.2	19.6	32.4	36.5
ROUGE	**20.7**	**33.6**	37.6	**19.8**	**32.6**	36.4
METEOR	18.2	33.4	**40.0**	17.2	32.5	**39.1**

Table 12: Reranking performance of the *BERT+S1:delex+S2* model variant with different scoring functions in Stage 2 for partitioning responses into the sets $\mathcal{R}_{high}$ and $\mathcal{R}_{low}$. For the models with ROUGE and METEOR as scoring functions, we perform model selection based on the best ROUGE and METEOR performance, respectively.

we only use the English partition of the whole dataset, which contains 2,952/295/442 dialogues for training/validation/testing. BiToD covers five domains: attraction, hotel, restaurant, weather, and

Context Size	Classification-based Reranking			Similarity-based Reranking		
	BLEU	ROUGE	METEOR	BLEU	ROUGE	METEOR
1	16.7	29.8	34.5	16.7	29.7	34.5
2	19.2	32.4	36.7	18.8	32.3	37.2
3	19.3	32.3	36.3	19.6	32.0	36.5
4	19.5	32.4	36.6	19.2	31.5	35.7
5	19.4	32.3	36.2	19.6	32.1	36.2

Table 13: Reranking performance of the *BERT+S1:delex+S2* variant with different context window sizes. The context window size is the number of historical dialogue utterances for the reranking models. We choose the same window size for fine-tuning at Stage 1, and note that the default context window size for the MinTL model is 3.

metro. We use the provided baseline system implemented with the mT5 model (Xue et al., 2021). We follow the default training script, only reducing the batch size due to hardware constraints (from 8 to 2).

We follow the setup from §4 for our reranking experiments. Following the baseline system, we only train our reranking model with lexicalised dialogues. **Table 7** lists the chosen and search set of our model hyperparameters. We searched the $k \in \{10, 100, 500, 1000, 5000\}$ for KNN regression and the number of anchors $n \in \{5000, 10000\}$. In addition, for BiToD we remove the downsampling step in Stage 2, which led to better empirical results.

E. Human Evaluation

We invited six human participants to join our human evaluation experiments.[11] Each annotator answered 100 questions indicating preferences over different model outputs. Those questions are formalised as A/B tests. Recently, comparative evaluation measurements have been shown to be more robust in human evaluation (Fomicheva et al., 2021). As demonstrated by **Figure 10**, given a dialogue context, the human evaluators are instructed to select better responses from pairs of responses generated from different models. We have compared among outputs from our classification-based models (quora-distilroberta), similarity-based models (BERT), and the greedy search baseline (best performing models in Table 2). The ordering of the tasks and models is completely randomised.

Table 14 shows full evaluation results. In addition, **Table 15** shows some example dialogues in the MultiWOZ test set, where the similarity-based reranked response is strongly preferred.

---

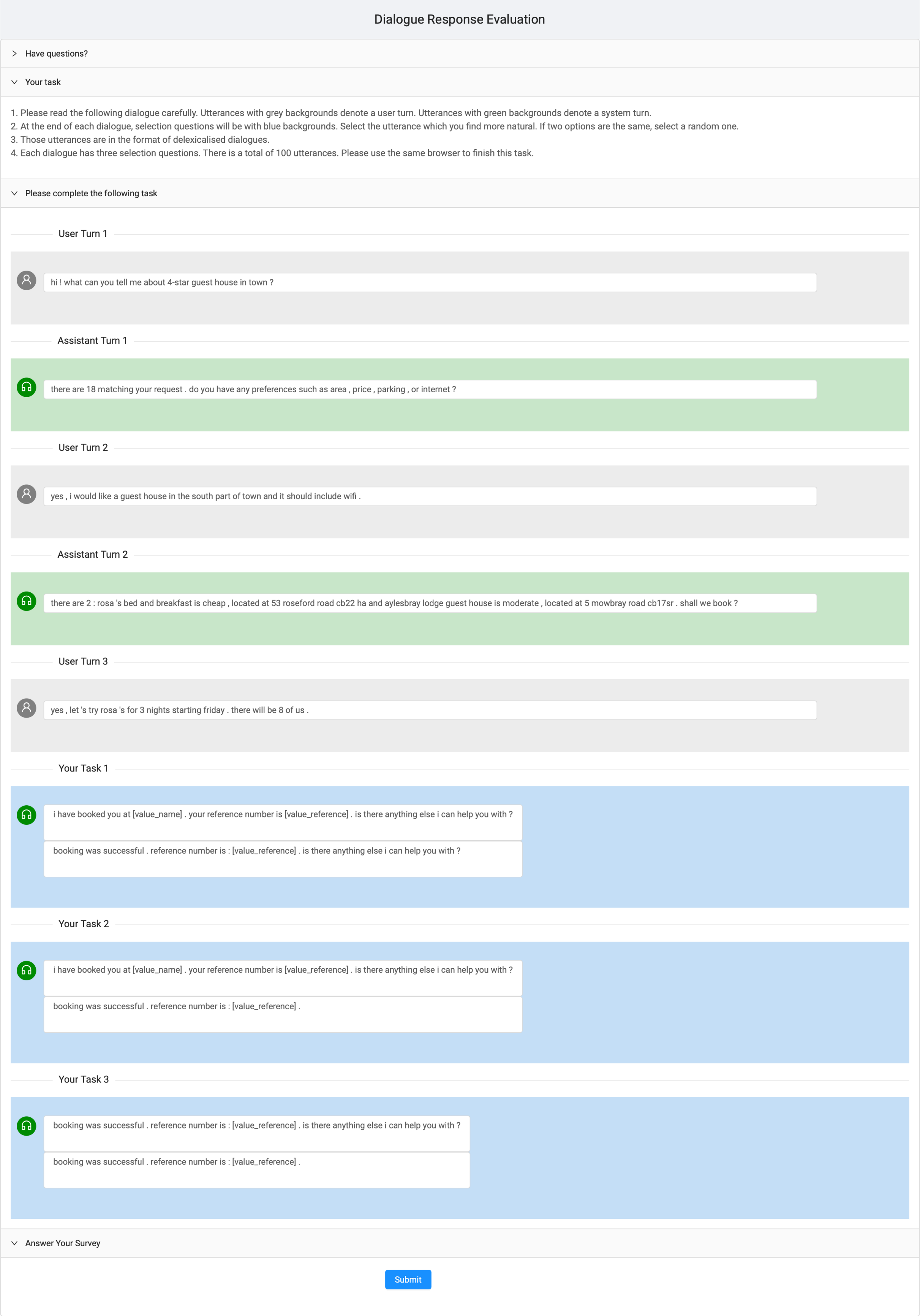
[11] The relevant departmental ethics committee has approved this experimental study. Our participants were recruited within the university who volunteered to join our experiment. They are students and academic staff who are proficient in English. The consent is obtained by signing a consent form.

Method A vs Method B	# of A	# of B	% of A	% of B	Total #	Fleiss' Kappa
Classification-based vs Greedy	297	303	49.5	50.5	600	0.28
Similarity-based vs Greedy	347	253	*57.8	*42.2	600	0.26
Similarity-based vs Classification-based	335	265	*55.8	*44.2	600	0.16

Table 14: Number (#) and percentage (%) of preferred responses during human A/B testing. Fleiss' Kappa is listed as a measure of inter-annotator agreement. * denotes statistical significance (2-sided binomial test, p < .05).

Example 1	
History	...... S: i would recommend christ 's college . it is free to visit . U: could i have their postcode please ? S: christ 's college 's post code is cb23bu U: thank you , that is all the information i need
Ground Truth	S: you are welcome , please contact us if we can help you with anything else .
Greedy	S: the postcode is [value_postcode] . is there anything else i can help you with ?
Classification-based	S: you are welcome . have a great day!
Similarity-based	S: ok, you are welcome . have a great day!
Example 2	
History	U: i am planning a trip in cambridge S: great ! is there a certain restaurant you are looking for ? U: moderate price in the east . and i need the phone number , address , and food type S: i have the curry prince which is indian food . would you like that 1 ? U: yes . that would be fine . what is the address and phone number ?
Ground Truth	S: the address is [value_address] and the phone number is [value_phone] . would you like me to make a reservation for you ?
Greedy	S: [value_name] is an [value_food] restaurant in the [value_area] . their address is [value_address] . their phone number is [value_phone] .
Classification-based	S: the address is [value_address] and the phone number is [value_phone] .
Similarity-based	S: the address is [value_address], and the phone number is [value_phone] .
Example 3	
History	...... U: thank you . can you also book a taxi for me ? S: sure , i need to know destination and where you need picked up . i also need the time and number of people . U: i also need a taxi from the church so i can get to the restaurant on time ."
Ground Truth	S: i will work on getting that booked for you .
Greedy	S: i can help with that . where will you be departing from ?
Classification-based	S: your taxi has been booked . it will be a [value_car] and the contact number is [value_phone] . is there anything else i can help you with ?
Similarity-based	S: your taxi is booked . it will be a [value_car] and the contact number is [value_phone] .
Example 4	
History	...... S: tr8259 will arrive in cambridge at 10:23 . would you like me to book a ticket for you on that train ? U: not yet . first i would like to get price and departure time . S: it will depart at 07:40 and it will cost 60.08 pounds . do you need help with anything else ? U: and i need a place to stay
Ground Truth	S: okay there are [value_choice] options . do you have a price preference or area
Greedy	S: i am sorry, but i don't have any [value_type] that meet your criteria . would you like to try a different price range or area ?
Classification-based	S: okay, what area would you like to stay in ?
Similarity-based	S: i can help you with that . what area of town are you wanting to stay in ?

Table 15: Selected example dialogues and corresponding responses where all five evaluators prefer the similarity-based reranked output compared to the greedy search output. S: system, U: user.



1. Please read the following dialogue carefully. Utterances with grey backgrounds denote a user turn. Utterances with green backgrounds denote a system turn.
2. At the end of each dialogue, selection questions will be with blue backgrounds. Select the utterance which you find more natural. If two options are the same, select a random one.
3. Those utterances are in the format of delexicalised dialogues.
4. Each dialogue has three selection questions. There is a total of 100 utterances. Please use the same browser to finish this task.

User Turn 1

hi ! what can you tell me about 4-star guest house in town ?

Assistant Turn 1

there are 18 matching your request . do you have any preferences such as area , price , parking , or internet ?

User Turn 2

yes , i would like a guest house in the south part of town and it should include wifi .

Assistant Turn 2

there are 2 : rosa 's bed and breakfast is cheap , located at 53 roseford road cb22 ha and aylesbray lodge guest house is moderate , located at 5 mowbray road cb17sr . shall we book ?

User Turn 3

yes , let 's try rosa 's for 3 nights starting friday . there will be 8 of us .

Your Task 1

i have booked you at [value_name] . your reference number is [value_reference] . is there anything else i can help you with ?

booking was successful . reference number is : [value_reference] . is there anything else i can help you with ?

Your Task 2

i have booked you at [value_name] . your reference number is [value_reference] . is there anything else i can help you with ?

booking was successful . reference number is : [value_reference] .

Your Task 3

booking was successful . reference number is : [value_reference] . is there anything else i can help you with ?

booking was successful . reference number is : [value_reference] .

Figure 10: Human evaluation: an example task for human-based A/B testing.