UNIVERSITY OF CAMBRIDGE

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

Songbo Hu[1], Ivan Vulić[1], Fangyu Liu[1], Anna Korhonen[1]

[1]Language Technology Lab, University of Cambridge, UK

LREC-Coling 2024 20 - 25 May 2024

Abstract

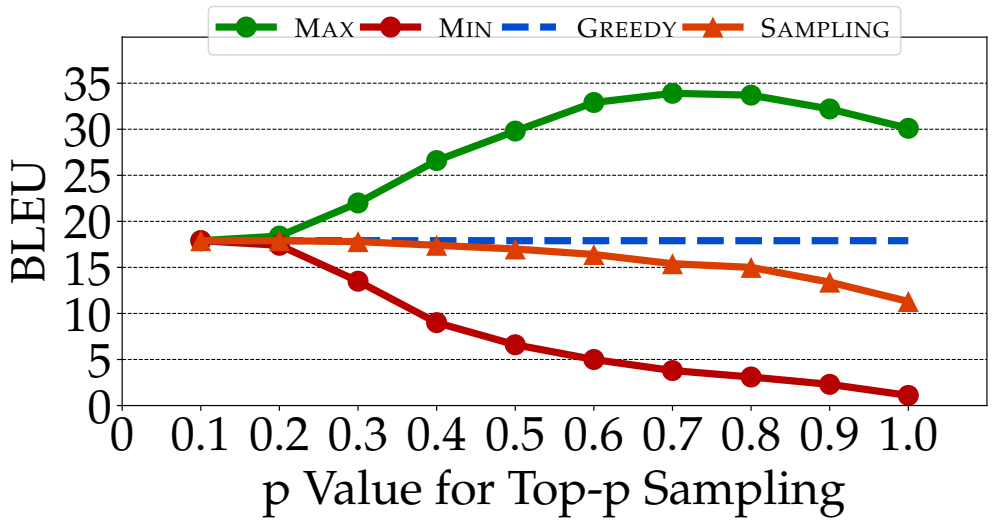
- Decoding from PLMs often encounters the 'likelihood trap' [1, 2], with standard strategies like greedy decoding, beam search, and truncated sampling facing inherent limitations.
- Overgenerating multiple responses exposes a broad diversity in response quality.
- We propose a simple but effective reranking method to select high-quality responses from initially overgenerated lists.
- Our method, validated on the MultiWOZ dataset, enhances a top E2E ToD system by 2.0 BLEU, 1.6 ROUGE, and 1.3 METEOR, setting new peak results.

Motivation: An Oracle Experiment

- In this ‘oracle’ experiment, we assume the availability of the ground truth response.
- We use the MinTL system [3] trained on the MultiWOZ 2.0 dataset [4].

Motivation: An Oracle Experiment

- We over-sampled 20 responses from this E2E ToD system with the nucleus sampling method [5].
- We ranked oversampled responses by their sentence-level BLEU similarity to the ground truth.



Preliminaries: Response Reranking

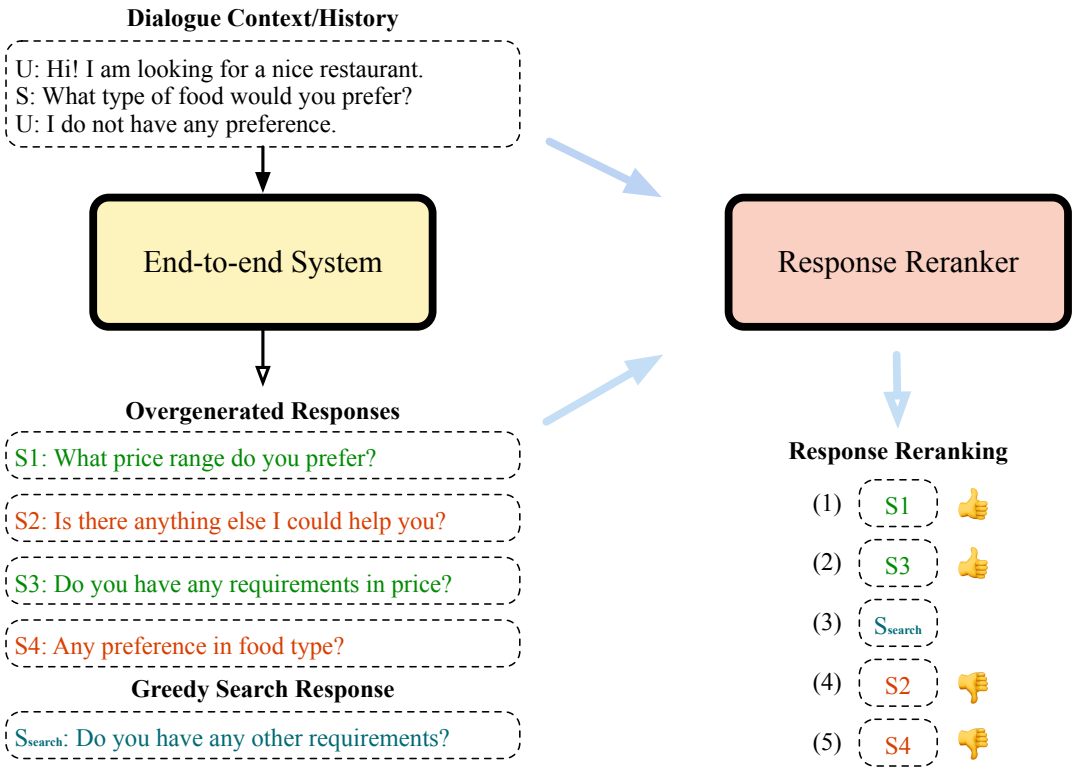
- Response reranking is similar to response selection.
- **Data:** $\mathcal{D}^{(i)}$ includes a dialogue context $c^{(i)}$ and its corresponding response $r^{(i)}$. The variables $\mathbf{c}^{(i)}$ and $\mathbf{r}^{(i)}$ represent their embeddings.
- **Training:** The goal is to develop a scoring function $s(\cdot, \cdot)$ that assigns a matching score to each context-response pair.
- **Inference:** Given a candidate set of responses R, the reranker evaluates each context-response pair using a scoring function $s(\cdot, \cdot)$ and selects the optimal response with $\text{argmax}_{r \in R}\, s\,(\mathbf{c}, \mathbf{r})$.
- **Objective:** Response reranking is tasked to improve the evaluation score $\text{M}(c, r)$.

Method in a Nutshell

- We train a generative E2E dialogue model $P_{\mathrm{MLE}}(r \mid c)$.
- 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 $P_{\mathrm{MLE}}(r \mid c)$, 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.
- The default scoring function is defined as the cosine similarity based on the `all-mpnet-v2` [6] encoder.

Method in a Nutshell

- After sampling responses $\mathcal{R}$, we categorise them into a high-scoring set $\mathcal{R}_{high}$ and a low-scoring set $\mathcal{R}_{low}$ based on their evaluation scores.

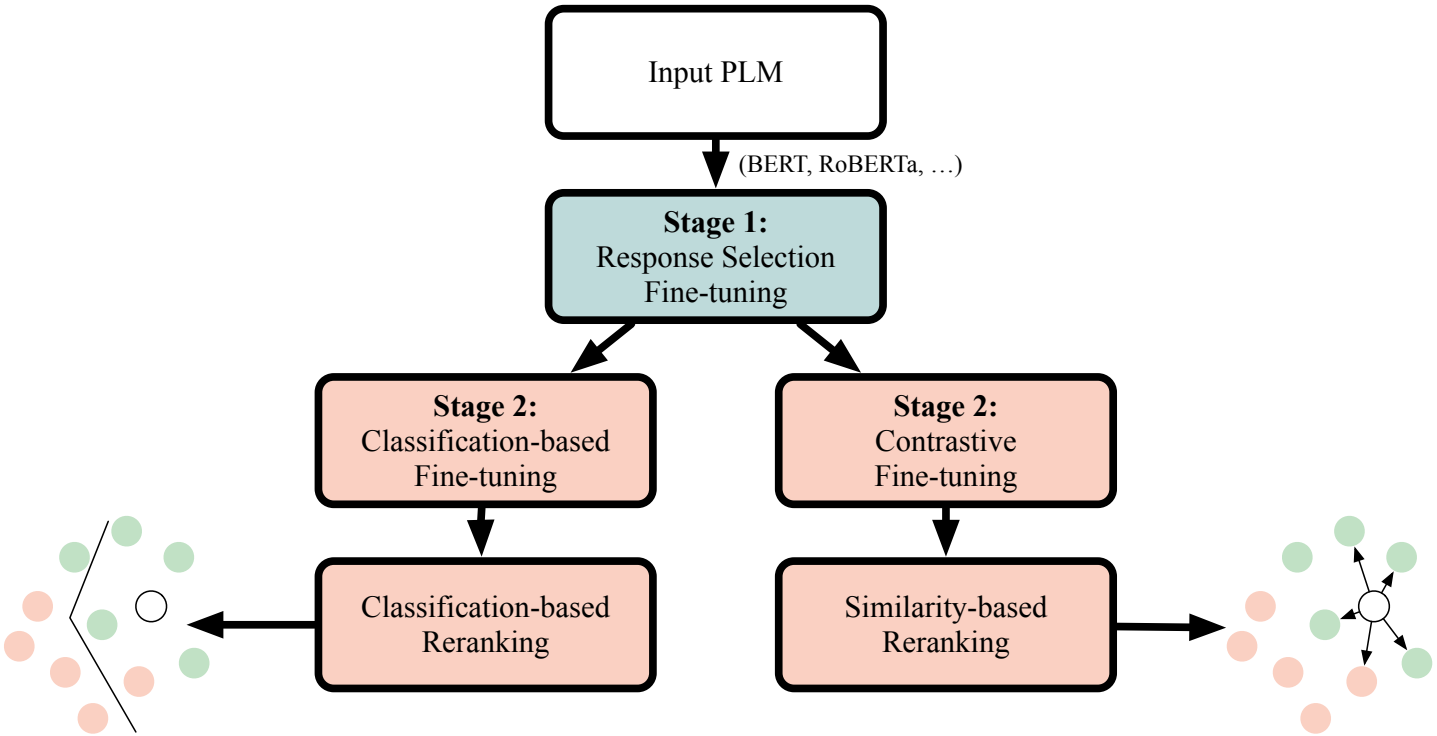


Method in a Nutshell

- 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 $P(r_k \in \mathcal{R}_{high} \mid c)$.
- Not require access to the ground truth during inference.

Method in a Nutshell

- 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.



Stage 1: Response Selection

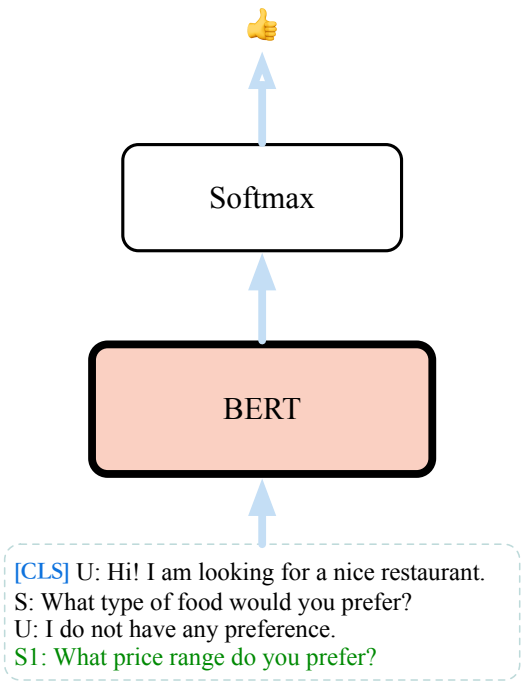
- We fine-tune PLMs on the response selection task.
- **Data:** For each dialogue pair $(c^{(i)}, r^{(i)})$ in dataset $\mathcal{D}$, we create a positive example $(c^{(i)}, r^{(i)}, 1)$ and N negative examples, e.g, $(c^{(i)}, r^{(j)}, 0)$.
- **Model:** We rely on a standard cross-encoder architecture. 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 $[c, r]$.
- **Binary classifier.**

Stage 2: Response Reranking

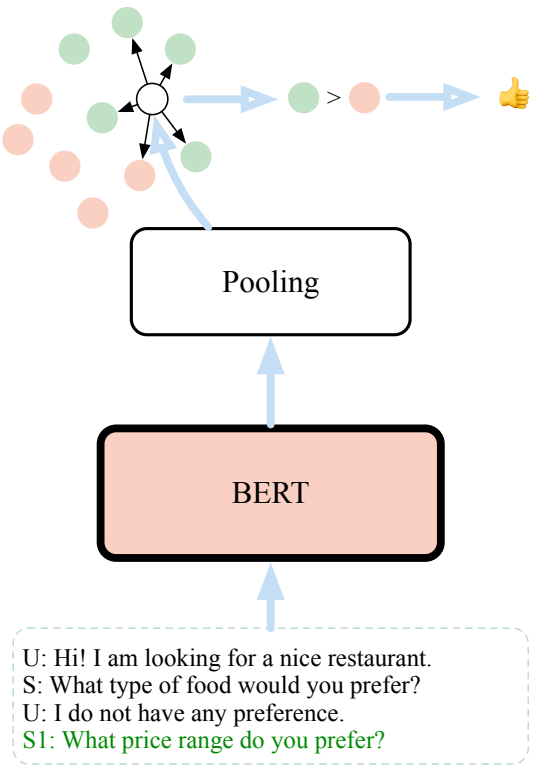
- Similarly, response reranking is again a binary classification task.
- **Data:** Each data entry is a tuple (c, r, l), where $l \in \{0, 1\}$. For each dialogue, we generate a set of responses $\mathcal{R} = \{r_1, r_2 \ldots r_j\}$ and a greedy search response r_{search}. The score s_{search} is used as a local *threshold value* that splits the set of generated responses into positive (i.e., 'high-quality') and negative ('low-quality').

Stage 2: Response Reranking

- **Method:** *classification-based* and *similarity-based*.



(a) Classification-based Reranking



(b) Similarity-based Reranking

Experimental Setup: ToD System

- **Dataset:** MultiWOZ 2.0 [4].
- **Baseline E2E System:** MinTL [3].
- **Decoding:** Greedy search and nucleus sampling [5] from the top-0.7 portion of the probability mass.
- **Evaluation Metrics:** Corpus BLEU [7], ROUGE-L [8], and METEOR [9] computed with delexicalised utterances.

Experimental Setup: Reranking Models

- **Input PLMs:** Sveral popular PLMs: BERT [10], RoBERTa [11], and their distilled versions [12]. Supervised sentence encoders: SimCSE [13] and other popular encoders from the sentence-transformers (i.e., SBERT) repository [6].
- **Model Variants:**
  - **PLM+S1+S2:** Stage 1 can be based on either lexicalised dialogues (*S1:lex*) or delexicalised (*S1:delex*) dialogues.
  - **PLM+S2**
  - **PLM+S1**
  - **PLM**
  - **Greedy**
  - **Sampling**

Results: Main Results

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

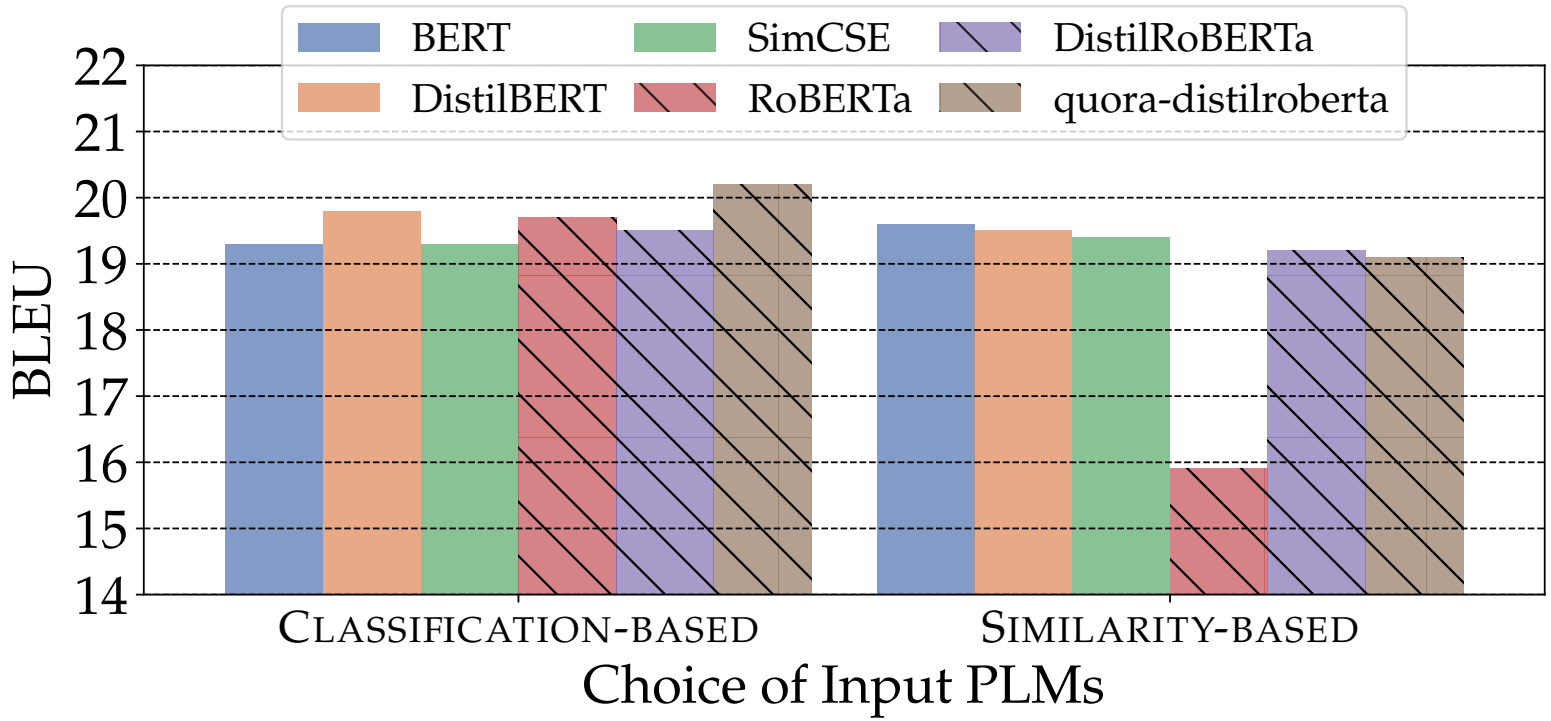
Results: Main Results

	Selection	Reranking		
Variant	R@1	BLEU	ROUGE	METEOR
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

Results: Ablation Study

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)

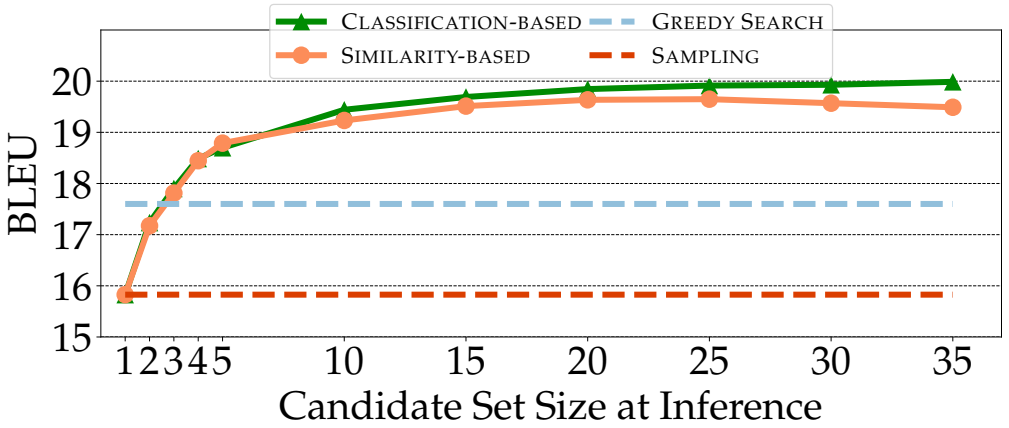
Results: Impact of the Input Encoder

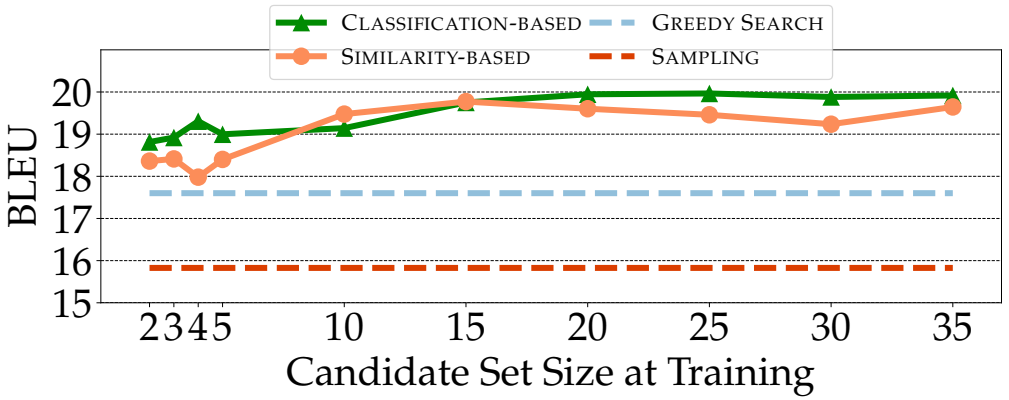


Results: Evaluation Metrics as Scoring Functions

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**

Results: Impact of the Candidate Set Size





Results: BiTOD

Variant	BLEU	ROUGE	METEOR
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

Results: Human Evaluation

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] .

Results: Human Evaluation

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

Thank you!

- We release the dataset and code at https://github.com/cambridgeltl/response_reranking.
- Any questions? Please email me: sh2091@cam.ac.uk.
- See you at LREC-Coling 2024!

References I

[1] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 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, November 2019. Association for Computational Linguistics.

[2] Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 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, April 2021. Association for Computational Linguistics.

References II

[3] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 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, November 2020. Association for Computational Linguistics.

References III

[4] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 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, 2018. Association for Computational Linguistics.

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

References IV

[6] Nils Reimers and Iryna Gurevych. 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, November 2019. Association for Computational Linguistics.

References V

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics.

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

References VI

[9] Satanjeev Banerjee and Alon Lavie. 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, June 2005. Association for Computational Linguistics.

References VII

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

References VIII

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

[13] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 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, November 2021. Association for Computational Linguistics.