Generate-and-Retrieve: use your predictions to improve retrieval for semantic parsing

Yury Zemlyanskiy[*†] **Michiel de Jong**[*†] **Joshua Ainslie**[‡] **Panupong Pasupat**[‡]
Peter Shaw[‡] **Linlu Qiu**[‡] **Sumit Sanghai**[‡] **Fei Sha**[‡]

[†] University of Southern California [‡] Google Research

{yury.zemlyanskiy,msdejong}@usc.edu

{jainslie,ppasupat,petershaw,linluqiu,sumitsanghai,fsha}@google.com

Abstract

A common recent approach to semantic parsing augments sequence-to-sequence models by retrieving and appending a set of training samples, called exemplars. The effectiveness of this recipe is limited by the ability to retrieve informative exemplars that help produce the correct parse, which is especially challenging in low-resource settings. Existing retrieval is commonly based on similarity of query and exemplar inputs. We propose GandR, a retrieval procedure that retrieves exemplars for which outputs are also similar. GandR first generates a preliminary prediction with input-based retrieval. Then, it retrieves exemplars with outputs similar to the preliminary prediction which are used to generate a final prediction. GandR sets the state of the art on multiple low-resource semantic parsing tasks.

1 Introduction

A common and successful approach to structured prediction problems (Li et al., 2021; Chen et al., 2020) is to treat the gold structure as a sequence and fine-tune a sequence-to-sequence model such as T5 (Raffel et al., 2020) or BART (Lewis et al., 2020). However, the performance of fine-tuned models suffers in low resource scenarios where available training data is limited relative to the complexity of the task (Chen et al., 2020).

Existing work (Pasupat et al., 2021; Gupta et al., 2021; Wang et al., 2022) has found that retrieving related training samples, denoted *exemplars*, and appending the retrieved input-output pairs to the sample input before processing the sample can improve performance in low resource settings. In principle, all information from exemplars is available to the model during training and could be stored in model parameters. However, in practice the model may not successfully retain all information, and reminding the model of salient input-output patterns at test time appears to help.

That raises the question: what exemplars are most informative for the model? Existing work focuses on retrieving exemplars for which the input is partially similar to the test input, effectively answering "*What is the output for similar inputs?*". In this work we explore whether there is complementary information in exemplars that answer the inverse question, "*What is the input for similar outputs?*".

We propose Generate-and-Retrieve (GandR), a method to retrieve exemplars with similar output as well as input. As the true output of a sample is in general unknown, GandR proceeds in two steps. First, a preliminary prediction is generated using retrievals with similar input only. Then, a new set of exemplars is retrieved based on a relevance measure that balances the similarity of the inputs and the similarity of the preliminary prediction and the exemplar output. Figure 1 provides an overview of the method.

We evaluate GandR in the setting of task-oriented semantic parsing, a core component of widely used virtual assistants. We show that similarity in output space provides a complementary signal to input similarity, yielding retrievals that prove more informative for the model. Moreover, for many structured prediction tasks the output space is more structured than the free-form input text, so that simple, non-learned distance measures work well for outputs even when inputs are lexically dissimilar. Table 5 demonstrates an example where our proposed similarity function retrieves an example that is somewhat less similarly phrased but with more similar output, and the model produces a better prediction as a result. Finally, the model has the opportunity to verify that its preliminary predictions are valid outputs in the target language.

The proposed method strongly improves performance in low-resource settings for semantic

[*] Work primarily done at Google Research.

Proceedings of the 29th International Conference on Computational Linguistics, pages 4946–4951
October 12–17, 2022.

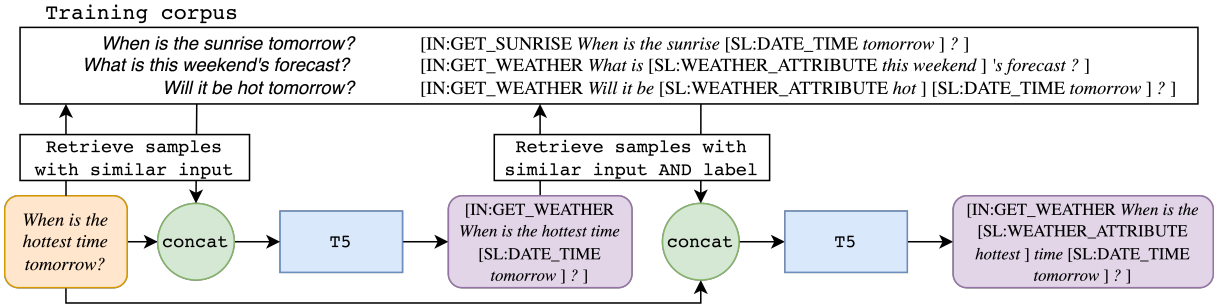

Figure 1: Overview of GandR. First, GandR generates a preliminary prediction using an input augmented with exemplars with similar inputs. Then, GandR retrieves exemplars based on a relevance measure balancing input similarity and similarity between the preliminary prediction and exemplar outputs, and generates a final prediction based on these exemplars.

parsing, achieving state of the art results for low-resource and transfer benchmarks in MTOP (Li et al., 2021) and TopV2 (Chen et al., 2020).

2 Method

We approach semantic parsing as a conditional language generation task and apply a T5 sequence-to-sequence model (Raffel et al., 2020) to predict a parse y given a query x. For each sample, we retrieve $K = 4$ relevant training exemplars sampled according to a relevance scoring function. We append the retrieved input-output pairs to the sample input and apply the T5 model to the augmented input to predict a parse output. In particular, let $(x'_1, y'_1), \ldots, (x'_K, y'_K)$ denote the retrieved input-output pairs, then the augmented input is

$$x' = x \;||\; x'_1 \;\&\; y'_1 \;||\; x'_2 \;\&\; y'_2 \;||\; \ldots$$

Our approach closely follows that of Pasupat et al. (2021), differing primarily in the choice of relevance function. During evaluation we retrieve the top K most relevant exemplars. During training, we sample retrievals according to a geometric distribution over the relevance score rank. In particular, the probability that we retrieve an exemplar is given by $p(1-p)^r$ where r is the rank of the exemplar according to relevance score and p is a temperature hyperparameter.

In Pasupat et al. (2021), the relevance score is given by the inner product of Universal Sentence Encoder (Cer et al., 2018) encodings of the candidate input and the sample input. We found that a simple TF-IDF (Ramos et al., 2003) similarity baseline achieves comparable or better results.

Our proposed approach, GandR, builds on the input-similarity baseline by constructing a hybrid similarity measure that takes into account not only the similarity between sample and candidate inputs, but also the similarity between the sequence predicted by the model and the candidate output. See Figure 1 for an overview. First, GandR generates a preliminary prediction using an input augmented with exemplars with similar inputs. Then, GandR retrieves exemplars based on a hybrid similarity measure over inputs and outputs, and generates a final prediction based on these exemplars.

Specifically, let $\hat{y}_i$ be preliminary prediction, then the proposed output similarity between samples i and j is given by the TF-IDF similarity between the predicted structure (in our case, the set of intents and slots) and the structure of the true parse y_j. Our proposed relevance score is a weighted sum of input and output similarity (see Appendix A for detailed description on how output similarity is computed):

$$R_{ij} = (1-\alpha)\text{TF-IDF}(x_i, x_j) + \alpha\text{TF-IDF}(\hat{y}_i, y_j)$$

2.1 Training

For simplicity, we train GandR in two stages. We start training with TF-IDF input relevance scoring, yielding model M_1. Model M_1 is used to generate GandR preliminary predictions during training and evaluation. We continue training M_1 for the remaining training steps, yielding M_2, which is used to generate final predictions augmented with retrievals from M_1. Note that this two-stage training is for convenience only, and it is possible to use a single set of weights M_{single} to generate preliminary and final GandR predictions. In that case, M_{single} needs to be trained with a mix of input-only and GandR retrieval augmentations to ensure it is able to use either effectively.

Model	$\text{MTOP}_{\text{boot}}$	MTOP_{1k}	$\text{MTOP}_{25\%}$	TOPv2_{W}	TOPv2_{R}
Reptile (Chen et al., 2020)				77.7	70.5
RAF (Shrivastava et al., 2022)				78.7	
CASPER (Pasupat et al., 2021)	73.3 / 83.9				
T5	72.9 / 83.3	62.8	78.5	79.2	68.8
T5 with input TF-IDF	74.9 / 84.5	67.2	79.4	79.9	71.0
GandR	**76.4 / 84.6**	**67.8**	**80.1**	**80.5**	**71.7**

Table 1: Results on semantic parsing benchmarks. We report the percentage exact match between true and predicted labels as sequences. Results are on test set for all benchmarks except $\text{MTOP}_{\text{boot}}$, where we report on dev to remain comparable with CASPER.

3 Related Work

Sequence-to-sequence models (Raffel et al., 2020; Lewis et al., 2020) have achieved state-of-the-art performance on task-oriented semantic parsing (Li et al., 2021; Chen et al., 2020; Aghajanyan et al., 2020) as well as other structured prediction tasks (Raffel et al., 2020). The general approach is to pre-train on language modeling and perform fine-tuning on the specific domain of interest.

Several works augment the input with retrieved exemplars from the training data, with differing methods for selecting informative examples. Pasupat et al. (2021) and Gupta et al. (2021) retrieve exemplars with similar input encodings from a pre-trained neural encoder, evaluating on semantic parsing. Wang et al. (2022) retrieves examplars for which the input has high BM25 similarity with the sample input, with good performance on language generation. We adopt a similar approach with TF-IDF similarity as a baseline for semantic parsing.

Black et al. (2021) and Das et al. (2021) learn dense retrievers in the spirit of Karpukhin et al. (2020), providing another path to incorporate label information for retrieval. Izacard et al. (2022) proposes other methods to fine-tune a dense retriever. These approaches require training a separate model specifically for retrieval, possibly with additional learning signal. In contrast, we employ a sparse similarity measure over model predictions that are produced incidentally in the course of fine-tuning the main model.

Selecting relevant training exemplars is also important for in-context prompting (Liu et al., 2021b). Similar to related fine-tuning literature, work in this direction uses either a pre-trained (Gao et al., 2020) or fine-tuned (Liu et al., 2021a) sentence encoder to retrieve exemplars.

Dataset	#Train	#Dev	#Test
MTOP	15667	2234	4385
MTOP_{1k}	1096	2234	4385
$\text{MTOP}_{25\%}$	3916	2234	4385
TOPv2_{S}	83703	11967	27336
TOPv2_{W}	176	147	5682
TOPv2_{R}	493	337	5767

Table 2: Dataset statistics: the number of examples per dataset and split.

Model	MTOP	TOPv2_{S}
RAF		87.1
CASPER	86.4	
T5	85.7	86.9
T5 input TF-IDF	86.4	87.0
GandR	86.4	87.0

Table 3: Performance on high-resource settings.

4 Experiments

4.1 Setup

We evaluate GandR and baselines on semantic parsing benchmarks MTOP (Li et al., 2021) and TOPv2 (Chen et al., 2020), focusing on low-resource and transfer settings. MTOP is a medium-sized semantic parsing dataset used in Pasupat et al. (2021), for which we evaluate on the *domain bootstrapping* setting in which one of the domains is limited to a very small amount of training data. We also evaluate on low-resource settings MTOP_{1k} and $\text{MTOP}_{25\%}$ in which we randomly sample 1k and 25% of training samples, respectively. TOPv2 is centered on transfer to low-resource domains: models are trained on a set of high resource-domains denoted as TOPv2_{S} and then fine-tuned on low-resource *Weather* and *Reminder* domains[1], denoted

[1]We are using 25 SPIS low resource split from Chen et al. (2020).

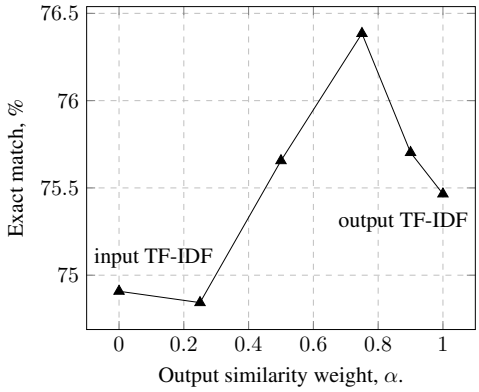

Figure 2: Performance on $\text{MTOP}_{\text{boot}}$ development set as a function of output similarity weight α.

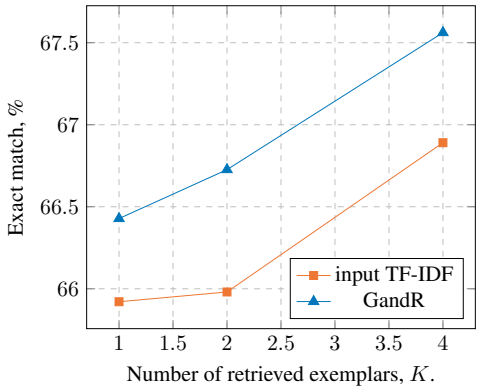

Figure 3: Performance on the development set of $\textbf{MTOP}_{\text{1k}}$ as a function of the number of retrieved exemplars K.

as TOPV2_{W} and TOPV2_{R}. We show the sizes of datasets and splits in Table 2. See Appendix B for details on the training setup.

4.2 Main results

The results of our primary experiments are shown in Table 1. We find that input TF-IDF is a strong baseline, rivaling or improving over prior work. Further, GandR retrieval outperforms all baselines, setting the state of the art on evaluated settings.

4.3 Ablations and discussion

Retrieval is less important for high-resource settings Table 3 shows results on the high-resource full MTOP and TOPv2 datasets. In higher-resource settings, augmenting the input with exemplars appears to be both less effective and less sensitive to retrieval method, with almost identical results among methods with and without retrieval for the highest resource TOPv2 dataset.

Retriever	$\text{MTOP}_{\text{boot}}$	TOPv2_{W}	TOPv2_{R}
input TF-IDF	35.9	55.1	20.1
output TF-IDF	70.3	74.8	53.7
GandR	70.0	68.7	52.5

Table 4: Template recall@K=4 on the development sets for $\text{MTOP}_{\text{boot}}$, TOPv2_{W} and TOPv2_{R}.

Using hybrid similarity leads to better retrieval quality Figure 2 displays $\text{MTOP}_{\text{boot}}$ performance as a function of TF-IDF output weight α. The results demonstrate that input and output similarity signals are strongly complementary. See Figure 4 and Figure 5 in the Appendix for similar experiments on the MTOP_{1k} and $\text{MTOP}_{\text{25\%}}$ benchmarks.

Considering output similarity leads to higher template recall Following Pasupat et al. (2021), we compute *template recall@K* as a proxy metric for retrieval. This measure corresponds to the proportion of evaluation samples for which at least one of the top K retrievals has the same template (identical intents and slots) as the gold parse. Results show (Table 4) that considering output as well as input similarity increases template recall. We note that output TF-IDF has similar or higher template recall than GandR even though it has lower performance. Ultimately, template recall is only a proxy, and we are really interested in retrieval *informativeness*; GandR's performance shows that balancing input similarity and template recall leads to exemplars that are most helpful for the model.

Hybrid similarity helps across different numbers of retrieved exemplars Figure 3 shows that GandR outperforms input TF-IDF similarity on MTOP_{1k} when we retrieve different number of exemplars: 1, 2 or 4. See Figure 6 in the Appendix for a similar experiment on the $\text{MTOP}_{\text{25\%}}$ benchmark.

4.4 Error analysis

The primary motivation for GandR is that hybrid similarity leads to more informative exemplars. Informativeness can only be objectively measured through model performance, but our motivating intuition appears to be borne out by samples in the data. We observe a number of different cases for which output or hybrid-similarity retrieval can help. Table 5 shows an example of a case for which input TF-IDF retrieves an irrelevant example with lexical overlap, while GandR retrieves an example with both lexical and parse overlap, leading to a

Input sample

x: *Could you connect me to the Musicals group*
y: [IN:CREATE_CALL [SL:GROUP *Musicals*]]

Training sample with similar input

x_1: *musicals in windham this weekend*
y_1: [IN:GET_EVENT [SL:CATEGORY_EVENT *musicals*] [SL:LOCATION *windham*] [SL:DATE_TIME *this weekend*]]
$\hat{y}$: [IN:CREATE_CALL [SL:CONTACT *me*] [SL:GROUP *Musicals*]]

Training sample with similar input and label

x_1: *can you please send text to the development group*
y_1: [IN:SEND_MESSAGE [SL:GROUP *development*]]
$\hat{y}$: [IN:CREATE_CALL [SL:GROUP *Musicals*]]

Table 5: Input TF-IDF retrieves an exemplar with lexical overlap ('musicals') that is not relevant to the sample. The GandR retrieval balances lexical and label similarity and leads to a correct prediction. Single representative exemplar out of 4 displayed for each method. See Table 9 in the Appendix for all retrieved exemplars.

correct prediction. Using preliminary predictions for retrieval can also allow the model to verify whether its predictions are correct. A common simple case when this can help is if the model generates a prediction that is dissimilar to any samples in the training set in which case the model may reconsider whether that prediction is correct (Table 7). Considering output similarity does come with tradeoffs. Table 8 demonstrates a situation where output similarity distracts the model away from a lexically similar and informative exemplar and the model is wrong as a result.

5 Conclusion

We propose GandR, a new method for structured prediction that generates a preliminary prediction, retrieves training exemplars with similar outputs (and similar inputs), and augments the input with the retrieved exemplars to generate a final prediction. We demonstrate that using output similarity yields improvements for semantic parsing in low-resource settings, achieving state of the art results on several semantic parsing benchmarks.

Acknowledgments

We thank William Cohen, Nicholas Fitzgerald and Luke Vilnis for insightful discussions and reviewers for their feedback. This work is partially supported by NSF Awards IIS-1513966/ 1632803/1833137, CCF-1139148, DARPA Awards#: FA8750-18-2-0117, FA8750-19-1-0504, DARPA-D3M - Award UCB-00009528, Google Research Awards, gifts from Facebook and Netflix, and ARO# W911NF-12-1-0241 and W911NF-15-1-0484.

References

Armen Aghajanyan, Jean Maillard, Akshat Shrivastava, Keith Diedrick, Michael Haeger, Haoran Li, Yashar Mehdad, Veselin Stoyanov, Anuj Kumar, Mike Lewis, and Sonal Gupta. 2020. Conversational semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5026–5035. Association for Computational Linguistics.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. 58.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta. 2020. Low-resource domain adaptation for compositional task-oriented semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5090–5100. Association for Computational Linguistics.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9594–9611. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Vivek Gupta, Akshat Shrivastava, Adithya Sagar, Armen Aghajanyan, and Denis Savenkov. 2021. RETRONLU: retrieval augmented task-oriented semantic parsing. *CoRR*, abs/2109.10410.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2950–2962. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Panupong Pasupat, Yuan Zhang, and Kelvin Guu. 2021. Controllable semantic parsing via retrieval augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7683–7698. Association for Computational Linguistics.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Akshat Shrivastava, Shrey Desai, Anchit Gupta, Ali Elkahky, Aleksandr Livshits, Alexander Zotov, and Ahmed Aly. 2022. Retrieve-and-fill for scenario-based task-oriented semantic parsing. *CoRR*, abs/2202.00901.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. *CoRR*, abs/2203.08773.