

Knowledge-grounded Dialog State Tracking

Dian Yu*, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent El Shafey, Hagen Soltau

Google Research

{dianyu, mingqiuwang, yuanc, izhak, shafey, soltau}@google.com

Abstract

Knowledge (including structured knowledge such as schema and ontology, and unstructured knowledge such as web corpus) is a critical part of dialog understanding, especially for unseen tasks and domains. Traditionally, such domain-specific knowledge is encoded implicitly into model parameters for the execution of downstream tasks, which makes training inefficient. In addition, such models are not easily transferable to new tasks with different schemas. In this work, we propose to perform dialog state tracking grounded on knowledge encoded externally. We query relevant knowledge of various forms based on the dialog context where such information can ground the prediction of dialog states. We demonstrate **superior performance** of our proposed method over strong baselines, especially in the few-shot learning setting.

1 Introduction

Pre-trained language models (LMs, Radford et al., 2019; Raffel et al., 2020) are the backbone of contemporary task-oriented dialog (TOD) models (Peng et al., 2020; Yang et al., 2021). However, the models are pre-trained on large generic corpora so that they do not contain **task-specific knowledge**. Previous work primarily suggests further pre-training or fine-tuning the LMs on in-domain data for adaptation (Wu et al., 2020; Hosseini-Asl et al., 2020), but it cannot consider information above the surface level. This makes it challenging for downstream tasks especially in the **few-shot learning setting** because mapping representation to the output space and encoding knowledge into the model parameters are entangled, while the latter may require more training data. Some more recent research proposes to incorporate external knowledge for response generation tasks (Dinan et al., 2019; Shuster et al., 2022; Chen et al., 2022;

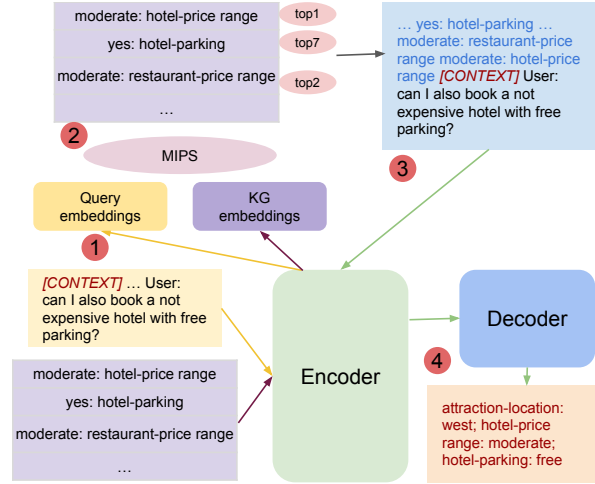


Figure 1: Model architecture for our proposed knowledge-grounded DST. The encoder first encodes the query and knowledge into representations, and we find the top-k most relevant knowledge elements to the context in step 2. We flatten the retrieved elements in step 3 and append to the query context as the input to the encoder-decoder model. The retrieved elements serve as a prior for DST.

Komeili et al., 2022), but it is not clear how to utilize such information for language understanding.

In TOD settings, because the API call structure is restricted to certain intents, slots, and values, the schema is often provided. For example, in a flight booking system, queries **like departure location and airlines are pre-defined**. Users, even though not bounded directly by what they can say to agents, have a limited and predictable vocabulary set to some extent. If the schema information is utilized, a model does not need to learn that “San Francisco” represents a departure place, rather than a general city name from the LM. This is particularly important for new information, such as movie titles or locations that do not appear in the LM training corpus. Similar to human annotators, grounding a dialog model on such knowledge makes it easier and more accurate in understanding conversations.

In this paper, we investigate knowledge-

*Work done while at University of California, Davis

grounded understanding for dialog state tracking (DST). In addition to using structured knowledge such as the ontology of slot type-value pairs, we also consider unstructured knowledge from the raw training data. We train a TOD model to query relevant knowledge for each turn in the context, and leverage the retrieved knowledge to predict dialog state. We evaluate our method on MultiWOZ (Budzianowski et al., 2018) for both the full-data and few-shot settings, and show superior performance compared to previous methods.

2 Related Work

2.1 Knowledge grounding

To relax the requirement of encoding knowledge of the whole world into model parameters, one direction is to disentangle knowledge representation from LMs. Most of these methods are applied to knowledge-intensive text generation tasks such as open-domain question answering (Lee et al., 2019; Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2021), and response generation with factual information (Dinan et al., 2019; Komeili et al., 2022; Thoppilan et al., 2022; Kim et al., 2020; Thulke et al., 2021; Chen et al., 2022). Similarly, some work also considers retrieving information to serve as a reference to refine the model generation process (Weston et al., 2018; Gonzalez et al., 2019; Khandelwal et al., 2021; Zhang et al., 2021). Different from these approaches, our method focuses on learning and utilizing available domain-relevant knowledge for language understanding tasks. Moreover, we propose to leverage knowledge of various formats.

2.2 Knowledge guided dialog understanding

Encoding domain schema into model parameters (Hosseini-Asl et al., 2020; Madotto et al., 2020) may not be efficient for unseen domains and tasks where the ontology can be different. One line of research (Ren et al., 2018; Wu et al., 2019; Zhou and Small, 2019; Rastogi et al., 2020; Du et al., 2021; Lee et al., 2021) leverages question-answering techniques to predict values for each slot, or prepend all slot-value information to the context (Zhao et al., 2022). However, this method is not scalable when the number of slot-value pairs is large, especially in multi-domain TOD systems. In addition, probably due to blurry attention over long context (Fan et al., 2021), Lee et al. (2021) find that adding potential slot values does not improve the model

performance. In contrast, retrieving only relevant schema effectively solves the scalability problem by specifying the knowledge with a fixed length.

Alternatively, instead of structured schema knowledge, recent research proposes to use hand-crafted demonstrations as prompts (Gupta et al., 2022) or find similar examples to guide understanding tasks (Yu et al., 2021; Pasupat et al., 2021; Yao et al., 2021) such as conversational semantic parsing. However, one turn can contain multiple dialog states so that retrieved examples from previous methods may not be sufficient to provide required evidence. Furthermore, our method can be applied to unify different forms of knowledge including structured and unstructured ones.

3 Methodology

Our proposed method is illustrated in Figure 1. Given the context \mathbf{x} , we first retrieve k relevant knowledge entries \mathbf{e} by the similarity between $\text{Enc}(\mathbf{x})$ and $\text{Enc}(\mathbf{e})$ using an encoder Enc . Then we integrate the retrieved entries e_1, e_2, \dots, e_k with the original context to form \mathbf{x}' , where \mathbf{x}' is used as the input for the target DST task.

Knowledge retrieval Different from previous work (such as question answering) where there is only one ground-truth knowledge for each query, multiple entries of the form slot-value pairs may exist in the ontology base that match the conversation context. Importantly, unlike passage retrieval where the query (e.g., a sentence) and the target (e.g., another sentence or passage) are similar to the pre-training corpus, structured knowledge such as schema pairs may have different representation distribution. Thus, an off-the-shelf encoder may retrieve noisy elements and degrade final performance, especially when training with the target task optimized on DST generation. Moreover, non-parametric retrieval methods such as TF-IDF and BM25 (Robertson and Zaragoza, 2009) rely on lexical overlapping, which could be detrimental when entries in schemas contain high word overlapping (e.g., same value for different slots).

We therefore train our knowledge retriever to promote similar representations between a query and its ground truth knowledge. We started with optimizing the marginal likelihood over all positive knowledge entries, but found that it resulted in peaky distribution centered around specific elements in our preliminary studies. Instead, we mini-

minimize binary cross-entropy with contrastive loss:

$$\mathcal{L} = - \sum_{i=1}^{i=n} y_i \cdot (\log(\text{sim}(\text{Enc}(\mathbf{x}), \text{Enc}(e_i))) + (1 - y_i) \cdot \log(1 - \text{sim}(\text{Enc}(\mathbf{x}), \text{Enc}(e_i)))) \quad (1)$$

where y_i is 1 if e_i appears in the target dialog state and otherwise 0. In our model, we use the same encoder Enc for both the context and the knowledge, and Enc is also used for the target DST task. sim defines the retrieval score, computed as the dot product between representations of the first token from the last layer¹.

Knowledge integration Once most relevant knowledge elements are retrieved by the model, this extra information can serve as a strong inductive bias to the downstream, knowledge-sensitive tasks. One common approach for knowledge integration is fusion-in-decoder (Izacard and Grave, 2021). Although efficient, it has been shown that retrieved information is likely to be ignored by a pre-trained model (Shuster et al., 2022). Hence, we concatenate retrieved knowledge with the context $\mathbf{x}' = e_k, e_{k-1}, \dots, e_1, \mathbf{x}$, where the entries are ordered from the least similar (e_k) to the most similar (e_1). The similarity can also be considered as the confidence an element e_i is relevant to the current context. We take the \mathbf{x}' as the context to the DST task. Therefore, our method is unified for knowledge of any format, and a bounded number of elements can solve the problem of memory constraint in previous research (Zhao et al., 2022).

4 Experiments and Results

4.1 Experiments and baselines

We conduct experiments on MultiWOZ 2.4 (Budzianowski et al., 2018; Ye et al., 2021) for DST in both full-shot and few-shot (1%) learning settings. For all experiments, we use T5-base and T5-XXL encoder-decoder models (Raffel et al., 2020) as the initial checkpoints. We use the publicly available T5 checkpoints² for our experiments. T5-base has 250 million parameters, and T5-XXL has 11B parameters. We train all models on 64 (for T5-base)

and 128 (for T5-XXL) TPU v3 chips (Jouppi et al., 2017). For fine-tuning, we set a learning rate of 1e-4 and a batch size of 32. We set the input and output sequence length to 1024 and 512 tokens. We train all models for 200k steps and report the performance on the test set from checkpoints achieving the best results on the development set. When multitask training on both retrieval and DST, we set 0.1 weight to the retrieval loss and 1 weight to the DST loss since it is relatively faster to converge for retrieval. We also experimented with 0.01, 0.05, 0.5, 1 for the retrieval loss weight, and found that 0.1 performs the best.

We compare with two baselines, seq2seq and D3ST. seq2seq takes the context as input, and predicts a sequence of linearized dialog state for each turn. Similarly, D3ST (Zhao et al., 2022) adds descriptions of each slot with potentially values as the prompt and predicts dialog states as multiple choice. Both baselines use the same T5 initial checkpoints. We report averaged joint goal accuracy (JGA) across three random seeds.

For our proposed method, we consider slot type (type), slot type and value (type+value), and training data (training) as knowledge sources. Specifically, for type, we consider all slot types (35 in total such as “hotel-parking”) as the knowledge base and retrieve corresponding top ten elements. For type+value, we consider each combination of types and their values in the form of “type: value” (1858 in total such as “hotel-parking: don’t care”) as the knowledge elements. Because there are more elements, we consider top 30 in our experiments for retrieval to achieve higher recall (with analysis later in Section 4.3). For training data, because of memory concerns, we randomly sample 500 training examples as the knowledge base and we consider the ground truth training example as the one with the highest F1 overlapping in the dialog slot types. We only consider top-1 due to the length constraint. For each knowledge source, we train retrieval together with the DST generator using the same model parameters.

4.2 Results

Table 1 shows DST results produced by different methods. Compared to the seq2seq baseline and D3ST, grounding on relevant knowledge by retrieval achieves better JGA by a large margin especially in the few-shot learning setting ($> 4\%$ absolute value). In the full data setting, our method

¹Other methods such as ColBERT (Khattab and Zaharia, 2020) are also applicable. In our preliminary experiments, we found that dot product is an effective measure, corroborating findings from Ni et al. (2021)

²https://github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md

		model	JGA	p	r
xxl	1%	baseline	50.24	-	-
		D3ST	54.37	-	-
		type	53.59	38.08	99.30
		type+value	55.32	12.94	48.78
		training	51.38	57.81	45.19
	full (100%)	baseline	73.18	-	-
		D3ST	75.90	-	-
		type	73.72	38.30	99.80
		type+value	75.47	12.49	68.56
		training	73.96	81.30	63.02
base	1%	baseline	30.48	-	-
		D3ST	16.37	-	-
		type	32.76	26.41	73.43
		type+value	34.89	6.17	21.95
	full (100%)	baseline	67.10	-	-
		D3ST	72.10	-	-
		type	70.51	37.64	98.80
		type+value	71.54	8.21	29.20

Table 1: Dialog state tracking results on MultiWOZ. We report averaged joint goal accuracy and retrieval metrics (precision and recall). With both T5-base and T5-xxl, grounding on retrieved slot types and values achieves better results by a large margin on the 1% few-shot learning setting, while performing on par with the full data setting.

performs on par with D3ST mostly due to that with more training data, the model can encode knowledge into parameters rather than relying on a separate, disentangled knowledge base. However, our method is not limited by the sequence length when we can specifically choose the number of retrieved elements regardless of the ontology size.

When comparing among different knowledge formats, type+value performs better than retrieving type only despite that retrieving is a harder task. As shown by recall³, with a large pre-trained model (XXL), recall for retrieving type only can achieve perfect scores (> 99%), but recall for type+value can only be 48% in the few-shot and close to 70% in the full-data setting. This indicates that the model can denoise distracting elements and make use of relevant knowledge as a positive inductive bias. Meanwhile, retrieving training data is similar to utilizing prompts (Gupta et al., 2022), but the worse performance compared to other knowledge formats suggests that selecting top-1 element is not optimal despite the relatively high recall. This is mostly due to that the retrieval results are noisy, as the small set of examples may contain slot types or values that are different from the ground truth. It

³Precision is determined by the number of retrieved elements we set, whereas recall measures the percentage of ground-truth knowledge elements being correctly retrieved. Therefore, recall is more informative.

is even less likely to find an example with exactly the same dialog state when the context is long. We leave further investigation by separating knowledge memory to support different knowledge sizes and external knowledge to future work.

4.3 Analysis

We study the relationship between retrieval and JGA in this section, and provide error analysis. We also analyze the detailed comparison between our method and D3ST.

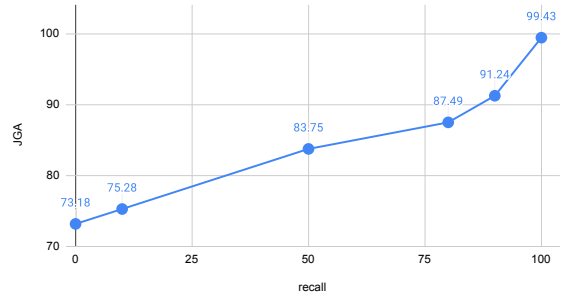


Figure 2: JGA with controlled retrieval recall from sanity check experimented with T5-XXL on the full-data setting. Results show that similar to our findings, even noisy retrieval improves model performance on DST.

Relationship between retrieval quality and JGA To understand the relationship between retrieval and the downstream task, we show JGA corresponding to recall in a controlled sanity check. Specifically, we randomly sample ground truth slot type-value pairs to match a target recall score and replace the rest dialog states with pairs uniformly sampled from the whole ontology (excluding the ground truth) without replacement as negative examples. Results (detailed in Figure 2) show that with T5-XXL on the full-data setting, 50% recall can significantly improve the model performance (83.75 JGA) while 90% recall can result in 91.24 JGA. This suggests that a high recall for retrieval is critical to JGA, while the model remains robust against noisy retrieval results. It also indicates that a better retrieval method (such as an external one Lazaridou et al. 2022) may achieve better performance. On the other hand, if we consider DST as a multi-class classification task with a retrieval module only, the model has to pick relevant elements from top-k, which is non-trivial.

We also consider separating retrieval from DST, i.e., train the model for retrieval first and then on

DST. Results show that although the model can achieve 97.38% recall, JGA actually drops to 70.33 on the full-data setting with XXL. We conjecture the main reason to be that different from freezing retrieval index in previous question-answering work, knowledge such as ontology or training data are more homogeneous and thus being more sensitive. This result is similar to our findings when training the two tasks jointly: retrieval metrics keep improving while JGA may drop with higher retrieval, even if we decrease the retrieval loss weight.

When we optimize separate parameters (i.e., two additional layers) for retrieval instead of the whole model, we observe slightly lower performance on JGA (54.76 compared to 55.32 on 1% data) and lower retrieval recall (36 compared to close to 46). Lastly, compared to top-30 with a JGA of 75.47, we observe an absolute drop of 0.40 for top-20, and 3.25 for top-10. This indicates that compared to noise in precision, retrieving ground-truth elements for recall is more critical to JGA.

Comparison to D3ST D3ST decodes the sequence of dialog state based on the order of slot types provided in the prompt by data pre-processing. In comparison, the order of retrieved elements varies while the order of dialog state depends on the ground-truth annotation. In other words, similar to the seq2seq baseline, our method requires learning the annotation order for DST prediction. This makes it more challenging to train, especially when there are similar knowledge elements retrieved. This can be justified by the slightly lower JGA with the full data setting. On the other hand, D3ST can be considered as a special setting of our grounding method where all knowledge elements are provided, and the DST generation model needs to implicitly detect relevant information and decode accordingly. We conjecture that the better performance on the few-shot setting over D3ST is due to that retrieving target elements while filtering noisy ones is easier than selecting corresponding knowledge, which can be shown from the high recall scores compared to lower JGA for D3ST. One future direction is to combine the benefits of the two worlds by utilizing the retrieved knowledge without length restriction.

Error analysis We found qualitatively that instead of ignoring retrieved elements as shown in previous research, the model does attend to retrieved slot-value pairs when decoding dialog states.

The main errors are from noisy retrieval, where a very similar elements with a higher rank (thus closer to the context in \mathbf{x}') than the ground truth knowledge may either stop the model from generating more states (i.e., missing target dialog states) or signal the model to generate the wrong elements directly. On the other hand, the model always predicts correctly if the ground truth are the most confident retrieved elements. To deal with the influence of attending only at the nearest few elements (which have the highest retrieval scores), we also experimented with randomly shuffling the retrieved knowledge but this results in lower scores (71.0 compared to 75.5) because the model needs to denoise from potential top-k elements without any additional information.

5 Conclusion

In this paper, we propose to disentangle domain knowledge and encode knowledge as a prior to dialog state tracking. Compared to previous research of grounding on knowledge for factual generation, our method can be applied to multiple sources of knowledge in the task-oriented dialog understanding setting. We conduct experiments on the MultiWOZ dataset and show superior performance especially in the few-shot learning setting. We plan to apply our method on more general natural language understanding tasks in the future.

6 Limitations

In the experiments, we show model improvements over strong baselines. Despite the simplicity of the method, we acknowledge that the domain ontology is not always available since knowledge (e.g., non-categorical slots) may not be a closed set, such as type+value in DST. However, this limitation can be lifted in two ways. Firstly, as shown in our experiments, retrieving slot type alone can also improve the model performance, which indicates that we may choose a knowledge base mixing type and type+value when the assumption that all values are predefined does not hold. Moreover, in most DST applications, the schema is specified before data collection and model training, where all target types and values need to match a database for information lookup. If the schema is unavailable, we may consider schema induction (Hudeček et al., 2021; Yu et al., 2022) where we can build the schema before DST. We plan to investigate these directions in our future work.

Acknowledgements

We thank Abhinav Rastogi from Google Research, and anonymous reviewers for their constructive suggestions.

References

- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. [Improving language models by retrieving from trillions of tokens](#). *CoRR*, abs/2112.04426.
- 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.
- Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadurai Sankar, Paul Crook, and William Yang Wang. 2022. [KETOD: Knowledge-enriched task-oriented dialogue](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. 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 *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Xinya Du, Luheng He, Qi Li, Dian Yu, Panupong Pasupat, and Yuan Zhang. 2021. [QA-driven zero-shot slot filling with weak supervision pretraining](#). 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 654–664, Online. Association for Computational Linguistics.
- Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. [Augmenting transformers with KNN-based composite memory for dialog](#). *Transactions of the Association for Computational Linguistics*, 9:82–99.
- Ana Valeria Gonzalez, Isabelle Augenstein, and Anders Søgaard. 2019. [Retrieval-based goal-oriented dialogue generation](#). Raghav Gupta, Harrison Lee, Jeffrey Zhao, Yuan Cao, Abhinav Rastogi, and Yonghui Wu. 2022. [Show, don’t tell: Demonstrations outperform descriptions for schema-guided task-oriented dialogue](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4541–4549, Seattle, United States. Association for Computational Linguistics.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. [Retrieval augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. [A simple language model for task-oriented dialogue](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191.
- Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. 2021. [Discovering dialogue slots with weak supervision](#). 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 2430–2442, Online. Association for Computational Linguistics.
- Gautier Izacard and Edouard Grave. 2021. [Distilling knowledge from reader to retriever for question answering](#). In *International Conference on Learning Representations*.
- Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. 2017. [In-datacenter performance analysis of a tensor processing unit](#). In *Proceedings of the 44th Annual International Symposium on Computer Architecture, ISCA ’17*, page

- 1–12, New York, NY, USA. Association for Computing Machinery.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. [Nearest neighbor machine translation](#). In *International Conference on Learning Representations*.
- Omar Khattab and Matei Zaharia. 2020. [ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT](#), page 39–48. Association for Computing Machinery, New York, NY, USA.
- 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.
- Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. [Internet-augmented dialogue generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.
- Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. [Internet-augmented language models through few-shot prompting for open-domain question answering](#).
- Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. [Dialogue state tracking with a language model using schema-driven prompting](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. [Latent retrieval for weakly supervised open domain question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. [Learning knowledge bases with parameters for task-oriented dialogue systems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2372–2394, Online. Association for Computational Linguistics.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. [Large dual encoders are generalizable retrievers](#). *CoRR*, abs/2112.07899.
- 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*, pages 7683–7698, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Building task bots at scale with transfer learning and machine teaching. *arXiv preprint arXiv:2005.05298*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *CoRR*.
- 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.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. [Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.
- Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. [Towards universal dialogue state tracking](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786, Brussels, Belgium. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. [Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion](#).
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze

- Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. [Lamda: Language models for dialog applications](#). *CoRR*, abs/2201.08239.
- David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. [Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog](#).
- 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.
- Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. [TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.
- Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. [Transferable multi-domain state generator for task-oriented dialogue systems](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.
- Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. [Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14230–14238.
- Xingcheng Yao, Yanan Zheng, Xiaocong Yang, and Zhilin Yang. 2021. Nlp from scratch without large-scale pretraining: A simple and efficient framework. In *ICML*.
- Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. [Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation](#). *CoRR*, abs/2104.00773.
- Dian Yu, Luheng He, Yuan Zhang, Xinya Du, Panupong Pasupat, and Qi Li. 2021. [Few-shot intent classification and slot filling with retrieved examples](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 734–749, Online. Association for Computational Linguistics.
- Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent Shafey, and Hagen Soltau. 2022. [Unsupervised slot schema induction for task-oriented dialog](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1193, Seattle, United States. Association for Computational Linguistics.
- Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2021. Joint retrieval and generation training for grounded text generation. *arXiv preprint arXiv:2105.06597*.
- Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022. [Description-driven task-oriented dialog modeling](#). *CoRR*, abs/2201.08904.
- Li Zhou and Kevin Small. 2019. [Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering](#). *CoRR*, abs/1911.06192.