RetroNLU: Retrieval Augmented Task-Oriented Semantic Parsing

Vivek Gupta[1,2*], **Akshat Shrivastava**[2], **Adithya Sagar**[2], **Armen Aghajanyan**[2], **Denis Savenkov**[2]
[1]School of Computing, University of Utah
[2] Facebook Conversational AI, Menlo Park
vgupta@cs.utah.edu ; {akshats, adithyasagar, armenag, denxx}@fb.com

Abstract

While large pre-trained language models accumulate a lot of knowledge in their parameters, it has been demonstrated that augmenting it with non-parametric retrieval-based memory has a number of benefits ranging from improved accuracy to data efficiency for knowledge-focused tasks such as question answering. In this work, we apply retrieval-based modeling ideas to the challenging complex task of multi-domain task-oriented semantic parsing for conversational assistants. Our technique, RetroNLU, extends a sequence-to-sequence model architecture with a retrieval component, which is used to retrieve existing similar samples and present them as an additional context to the model. In particular, we analyze two settings, where we augment an input with (a) retrieved nearest neighbor utterances (utterance-nn), and (b) ground-truth semantic parses of nearest neighbor utterances (semparse-nn). Our technique outperforms the baseline method by 1.5% absolute macro-F1, especially at the low resource setting, matching the baseline model accuracy with only 40% of the complete data. Furthermore, we analyse the quality, model sensitivity, and performance of the nearest neighbor retrieval component's for semantic parses of varied utterance complexity.

1 Introduction

Roberts et al. (2020) demonstrated that neural language models quite effectively store factual knowledge in their parameters without any external information source. However, such implicit knowledge is hard to update, i.e. remove certain information (Bourtoule et al., 2021), change or add new data and labels. Additionally, parametric knowledge may perform worse for less frequent facts, which don't appear often in the training set, and "hallucinate" responses. On the other hand, memory-augmented models (Sukhbaatar et al., 2015) decouple knowledge source and task-specific "business logic", which allows updating memory index directly without model retraining. Recent studies showed their potential for knowledge-intensive NLP tasks, such as question answering (Khandelwal et al., 2020; Lewis et al., 2020c).

In this work, we explore RetroNLU: retrieval-based modeling approach for task-oriented semantic parsing problem, where explicit memory provides examples of semantic parses, which model needs to learn to transfer to a given input utterance. An example semantic parse for task-oriented dialog utterance and its corresponding hierarchical representation are presented in Figure 1.

Utterance: Driving directions to the Eagles game

Semantic Parse: [IN:GET_DIRECTIONS Driving directions to [SL:DESTINATION [IN:GET_EVENT the [SL:NAME_EVENT Eagles] [SL:CAT_EVENT game]]]]

Tree Representation:

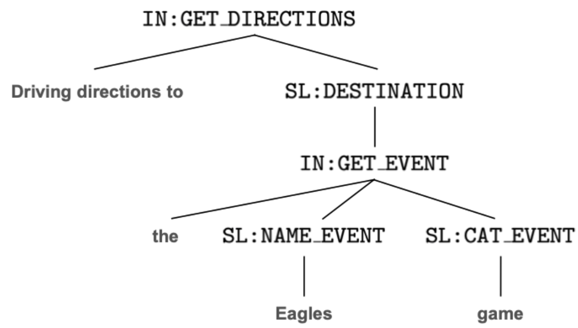

Figure 1: An intent-slot based compositional semantic parsing example(coupled) from TOPv2 (Chen et al., 2020).

In this paper we are focusing on the following questions: (a) *Data Efficiency*: Can retrieval based on non-parametric external knowledge alleviate reliance on parametric knowledge typically acquired via supervised training on large labeled datasets?[1] We examine how different training settings, depending on the amount of supervision data available,

*Work done by author while interning at Facebook Conversational AI.

[1]Parametric knowledge is information stored in model parameters. Non-parametric knowledge refers to external data sources that the model uses to infer.

impact model prediction, i.e. fully supervised vs. limited supervised training. (b) *Semi-supervised Setting*: Can we enhance models by using abundant and inexpensive unlabeled external non-parametric knowledge rather than structurally labeled knowledge? We examine the effect of utilizing unlabeled similar utterances instead of labelled semantic parses as external non-parametric knowledge on model performance. (c) *Robustness to Noise*: Can a model opt to employ parametric knowledge rather than non-parametric knowledge in a resilient manner, for example, when the non-parametric information is unreliable? We examine the model's resilience and its reliance on non-parametric external information. External knowledge is not always precisely labeled and reliable for all examples/utterances. (d) *Utterance Complexity*: Is non-parametric external knowledge addition effective for both uncommon and complex structured (hierarchical) examples? We examine whether external knowledge addition is more beneficial in certain cases than others, or if it supports accurate predictions for all situations equally. It would be fascinating to investigate if external information could also help enhance difficult and complex examples/utterances. Finally, we examine the upper limit on the utility of external information. We examine structural redundancy concerns in nearest neighbor retrieval. (e) *Knowledge Efficiency*: Is it beneficial to continue adding external information, or are there certain boundaries and challenges? Our contribution are as follows:

1. We demonstrate that combining parametric and non-parametric knowledge enhance model performance on the complex structured task of task-oriented semantic parsing.

2. We illustrate the effectiveness of our approach in a critical situation of learning with sparse labeled data (i.e. limited parametric knowledge).

3. We establish the efficacy of retrieval-based method in semi-supervised settings, where model's input is supplemented with unannotated instances (i.e. unlabeled examples).

4. By comparing predictions on clean vs. noisy neighbours, we establish the model's resilience to external non-parametric knowledge quality.

5. Finally, we examine performance gains with inputs of varying complexity: semantic structure composition and it's frequency (i.e. frequent/rare).

Overall, we demonstrate that retrieval enhanced method can improve performance on complicated structured prediction tasks like task oriented semantic parsing without extra supervision. Furthermore, the augmentation approach is data efficient and performs well in low resource settings with limited label data. The dataset, and associated scripts, will be available at `https://retronlu.github.io`.

2 Proposed Approach

Our proposed approach consists of four main steps: (a) **index construction** by embedding training examples and computing cosine similarity; (b) **retrieval**, where we extract the nearest neighbor utterances from the index given an example utterance; (c) **augmentation**, in which we append the nearest-neighbor utterance ground truth semantic parse (semparse-nn) or the utterance itself (utterance-nn) to the original input via a special separator token (such as '|'); and (d) **semantic parsing**, in which we train the parsing model using the retrieval-augmented input with output ground truth. Figure 2 illustrates the Retrieval Augmented Semantic Parsing (RETRONLU) approach.

Indexing: To build an index we use a pre-trained BART model to get training utterance embeddings. More specifically, we get sentence embedding for all the training utterances. These sentence embeddings are obtained as average of token embeddings from last model layers of the BART models.[2] We then used the cosine similarity between embeddings to build a fast and efficient retrieval index with efficient FAISS library (Johnson et al., 2019).

Retrieval: Next, given a new input (training or test row), we obtain embeddings by running it through same pre-trained BART, and then query the index with it to retrieve nearest neighbors text and their ground truth semantic parses based on cosine similarity. For training data, we exclude an example itself from the retrieved list. For example, for input utterance *"please add 20 minutes on the lasagna timer"*, we retrieve the nearest neighbour *"add ten minutes to the oven timer"* along

[2]extract_features function `https://github.com/pytorch/fairseq/tree/main/examples/bart`

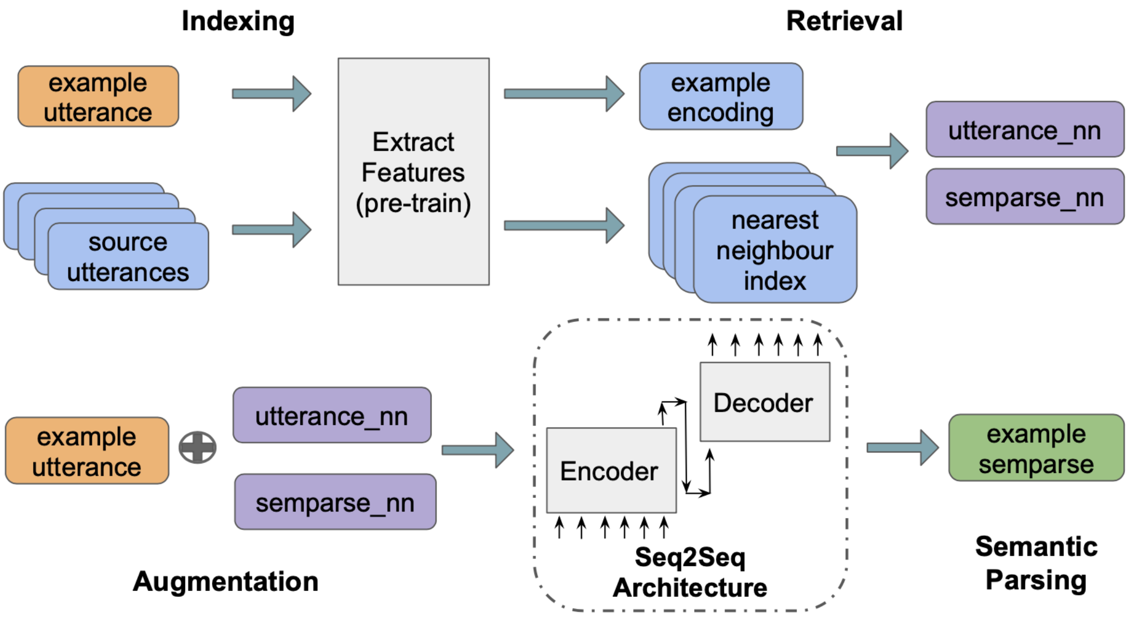

Figure 2: High level flowchart for retrieval augmented semantic parsing (RETRONLU) approach.

with the semantic parse as *"[in:add_ time_ timer add [sl:date_ time ten minutes] to the [sl:timer_ name oven] [sl:method_ timer timer]]"*.

Augmentation: Once we got a list of nearest neighbors, we can append either utterance text or semantic parse to the input, following the left to right order.[3] The closest neighbor appears to the immediate left of the input example utterance. One can also directly append the nearest neighbor utterance rather than the semparse, refer as utterance-nn. For the last example the final input would after augmentation would be *"[in:add_ time_ timer add [sl:date_ time ten minutes] to the [sl:timer_ name oven] [sl:method_ timer timer]] | please add 20 minutes on the lasagna timer"* for semparse-nn, and *"add ten minutes to the oven timer | please add 20 minutes on the lasagna timer"* for utterance-nn. Here, the token ' | ' act as a separator between the input utterance and the neighbour's.

Semantic Parsing: The final step is to train a sequence-to-sequence model such as LSTM or Transformer. We fine-tune a BART model with copy mechanism (Aghajanyan et al., 2020), which incorporates benefits of pre-trained language model (BART) and sequence copy mechanism (copy-ptr), and most importantly obtain state-of-the-art results on the TOPv2 (Chen et al., 2020), a challenging task oriented semantic parsing dataset with hierarchical compositional instances. The retrieval augmented example is an input to the encoder and the corresponding ground-truth semantic parse as the labeled decoded sequence. At test time, we simply pass the augmented input to the trained RETRONLU model, and take it's output as the predicted semantic parse for the input utterance.

[3] We followed GPT-3 and other generation model, where task examples are pre-pended to the input. Hence, utterance is always nearest to the decoder followed by the first nearest neighbour inorder.

3 Experiment and Analysis

Our experiments examines how our knowledge retrieval-based augmentation technique impacts model performance indicators such as accuracy and data efficiency. We study the following questions:

RQ1. Can today's pre-trained models leverage non-parametric information in manner as described in §2 to enhance task-oriented semantic parsing?

RQ2. If only part of the dataset has semantic parses, i.e. limited supervision setting, can augmentation with unannotated instances (utterance_nn) enhance semantic parsing accuracy?

RQ3. How efficient is a retrieval-augmented model in terms of data? Is it more accurate even with less training data than the baseline seq2seq model?

RQ4. Does non-parametric memory benefit instances equally, e.g., do we notice greater benefits for (a) more complex (i.e. compositional) or (b) less frequent semantic frames (i.e. tail over head)?

RQ5. (a) Does augmentation with more nearest neighbors benefits? (b) How sensitive is the model to retrieval noise? Can the model predict right intent/slots for low-quality retrieve instances?

Our experiments are designed to demonstrate how non-parametric external information can be beneficial to a parametric model and to undertake an in-depth assessment of the impact.[4]

3.1 Experimental setup

In this section, we discuss the datasets, pre-processing, and the model used in the experiments.

Datasets. For our experiments, we used the multi-domain complex compositional queries based popular TOPv2 (Chen et al., 2020) dataset for task-oriented semantic parsing. We concentrated our efforts on task-oriented parsing because of the commercial importance of data efficiency requirements in conversational AI assistants dialogues.[5] The TOPv2 dataset contains utterances and their semantic representations for 8 domains: source domains - 'alarm', 'messaging', 'music', 'navigation', 'timer', and 'event', and target domains: 'reminder' and 'weather', designed especially to test the zero-shot setting. For our experiments we chose source domains, which has a good mixture of simple (flat) and complex (compositional) semantic frames. For dataset statistics refer Table 1 in Chen et al. (2020).

Data Processing. To build a retrieval index we used the training split of the dataset. Each utterance was represented by its BART-based embedding and indexed using FAISS library (Johnson et al., 2019). [6] With FAISS computation cost of updating indexing was kept to bare minimum. The only additional cost will be increase in inference time due to augmented neighbor. To produce augmented examples, we retrieved nearest neighbors for each training and test examples from the training set, except excluding all training instances with exact utterance matches. In the augmented examples, we use the special token '|' to separate the nearest neighbors, as well as utterance with the first neighbor.[7] We used only one neighbor for most experiments except when we analyse multiple neighbors effects on performance.

In nearest neighbor augmented input, we followed right to left order, where the actual model input comes last, and its highest ranked neighbor is appended to the left of the utterance, followed by other neighbors in the left based on their ranking. [8] For input data pre-processing, we follow (Chen et al., 2020) procedure, we obtain BPE tokens of all tokens, except ontology tokens (intents and slot labels), which are treated as atomic tokens and appended to the BPE vocabulary. Furthermore, we use the decoupled canonical form of semparse for all our experiments. For decoupling, phrases irrelevant to slot values are removed from semparse, and for canonicality, slots are arranged in alphabetic order (Aghajanyan et al., 2020).

Models. For fair comparison with the earlier baseline, we use the state-of-the-art BART based Seq2Seq-CopyPtr model for task-oriented semantic parsing. [9] The BART based Seq2Seq-CopyPtr model initialize both the encoder and decoder with pre-trained BART (Lewis et al., 2020b) model and also use the copy mechanism similar to See et al. (2017), refer Chen et al. (2020) for details. We choose the BART based Seq2Seq-CopyPtr model for the task because it's a strong baseline, the performance of the other language model such as RoBERTa without augmentation was inferior (Chen et al., 2020; Aghajanyan et al., 2020). On out-of-domain instances, RoBERTa-CopyPtr performs 0.6 % worse than BART-CopyPtr.[10] The model is using the copy mechanism (See et al., 2017), which enables it to directly copy tokens from the input utterance (or from example semantic parses from nearest neighbors).

Hyperparameters. We use the same default hyper-parameters for all models training , i.e. baseline (without-nn) and RETRONLU models (utternce-nn, semparse-nn). For training we use 100 epochs, Adam optimizer (Kingma and Ba, 2014) with learning rate α of $1e-4$ and decay rate β_1 and β_2 of 0.9 and 0.98 respectively in all our experiments. Also, we didn't added any left or right padding and rely on variable length encoding in our experiments. We use warm-up steps of 4000,

[4] We did not seek to modify the architecture which ensure the augmentation methodology is flexible.

[5] Regardless of augmented neighbors structure the approach remain consistent.

[6] We use L2 over unit norm BART embedding for indexing.

[7] Using different separator tokens for neighbor-neighbor pair and utterance-neighbor pair didn't improve performance.

[8] Similar performance is obtained by ordering utterances left to right, followed by their neighbors in index order.

[9] We prefer transformer-based language model over non-transformer models, such as LSTM, because the later does not capture extended context as well as the former.

[10] Our findings, however, we believe, are universal and can be applied to different models, including RoBERTa.

dropout ratio of 0.4, and weight decay 0.0001, but no clip normalization as regularization during the training. We use batch size of 128 and maximum token size of 2048. Furthermore, to ensure both encoder and decoder BART, can utilise the extra nearest neighbour information, we increase the embedding dimension to 1024.

3.2 Results and Analysis

This section summarizes our findings in relation to the aforementioned research questions.

Full Training Setting. To answer RQ1, we compare performance of original baseline (without-nn) with our retrieval augmented models, i.e. augmenting first neighbour utterance (utterance-nn) and augmenting first neighbour semantic parse (semparse-nn). Table 1 compares the frame accuracy of retrieval augmented (a) top nearest neighbour utterance (utterance-nn), (b) top nearest neighbour ground-truth semantic parse (semparse-nn) with original baseline (without-nn) with model train on complete training data.

Domains	without-nn	utterance-nn	semparse-nn
Alarm	86.67	87.17	**88.57**
Event	83.83	**85.03**	84.77
Music	79.80	**80.73**	**80.71**
Timer	81.21	**81.75**	81.01
Messaging	93.50	94.52	**94.65**
Navigation	82.96	84.16	**85.20**
micro-avg	84.43	85.28	**85.74**
macro-avg	84.66	85.56	**85.82**

Table 1: Performance of RETRONLU w.r.t original baseline (without-nn) with full training.

Analysis: We observe performance improvements with retrieval-augmented models for most domains compared to the original baseline (without-nn) in both cases. The increase in performance (micro-avg) is more substantial 1.4% with semparse-nn compare to 0.85% with utterance-nn. The improvement in utterance-nn augmentation performance is likely due to memorization-based generalization, as explained earlier by (Khandelwal et al., 2019).[11] The results shows the retrieval augmented semantic parsing is overall effective. Furthermore, the performance enhancement can be obtained also with unstructured utterance (utternace-nn) as nearest neighbour. The utterance-nn based augmentation is particularly beneficial in semi-supervised scenarios, where we have a large unlabelled dataset.

[11]The scores are averaged over three runs with std. of 0.3%

Limited Training Setting. To answer RQ2, we compare model performance which are trained with limited training data. Figure 3 shows frame accuracy (micro-avg) when we use only 10% to 50% of the training data. The training datasets are created in an incremental setting so that next set include examples from the former set. Additionally, we use the complete index to retrieve the nearest neighbors.

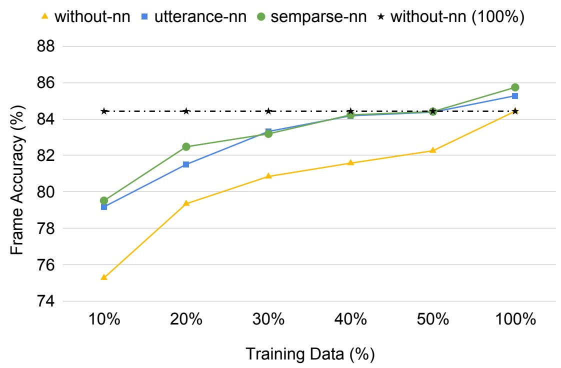

Figure 3: Performance of RETRONLU w.r.t original baseline (without-nn) with limited supervised training. The x-axis is linearly scaled upto 50% data.

Analysis: As expected, the performance of all models increases with training set size. Both retrieval augmented models i.e. utterance-nn and semparse-nn outperform the without-nn baseline for all the training sizes. The improvement via augmentation is more substantial with less training data, i.e. 4.24% at 10% data vs 1.30% at 100% data. Furthermore, the semparse-nn augmented model outperforms the original completely train (without-nn) model with only 40% of the data (i..e RQ3). The results show that the retrieval augmented semantic parsing is more data efficient, i.e. when there is (a) limited labelled training dataset with more unlabelled data for indexing (utterance-nn), and (b) sufficient training data but limited training time (semparse-nn).

The first case is useful when the ground truth label is missing for utterances due to lack of annotation resources. In such a scenario, one can build the index using large amount of unlabeled utterances and use the index for augmentation. The second case helps us train the model faster, while maintaining all annotated examples in the index. In such a case, one can update the retrieval index only, without re-training the model again and again. This is useful when training on full data is not possible due to limited access to model (black-box), a cap on the computation resources, or for saving training time i.e. industries fast deployment need. E.g. There is a constant stream of bugs relating to misclassified examples in production systems. Our RETRONLU approach enables rapid adjustment of the system's behavior without retraining or establishing a new model.

Effect of Utterance Complexity. To answer RQ4(a), we analyse the retrieval augmented model performance improvements (with full training) on simple utterance with only one level in semantic representation (depth-1) vs complex utterance with hierarchical semantic frames (compositional depth-2 and above). Figure 4 shows frame accuracy of without-nn, utterance-nn and semparse-nn model with utterance complexity.

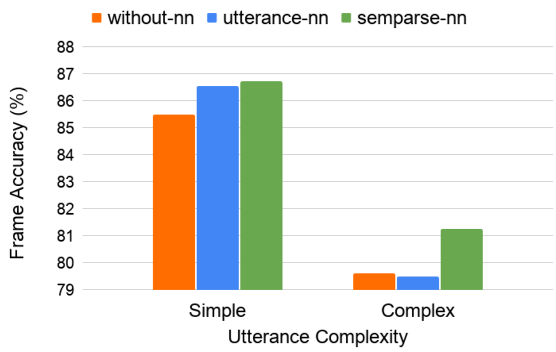

Figure 4: Performance comparison (micro-avg) of RETRONLU w.r.t original baseline (without-nn) with utterance complexity, i.e. simple and complex.

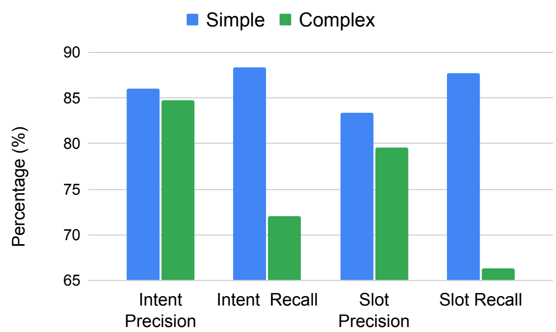

Figure 5: Precision and Recall of intents and slots for semparse-nn nearest neighbour w.r.t to gold semparse.

Analysis: As expected, all models perform relatively poorly on complex utterances (79.5%) in comparison to simple utterances (85.5%). Interestingly, both augmentation models equally improve performance on simple queries. And with semantic-frame based augmentation we observe a substantial performance improvement on complex challenging utterances, of 2%, with respect to the original baseline (without-nn). This suggests, that by retrieving nearest neighbors and providing a model with examples of complex parses, the model learns to apply it to a new request. Figure 5 shows precision and recall for intents and slots in retrieved semantic parses. The recall for intent and slot retrieval is 15% lower for complex utterances. [12] Thus, highlighting one reason for a performance gap between simple and complex frames.

Effect of Frame Rareness. To answer RQ4(b), we analyze the retrieval augmented model performance improvement (with full training data) with frame rareness, as shown in Figure 6.

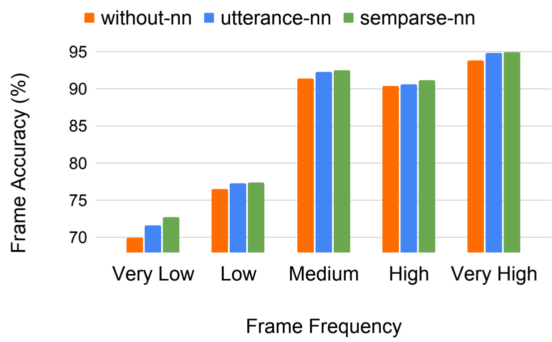

Figure 6: Performance of RETRONLU w.r.t original baseline (without-nn) with varying frame frequency.

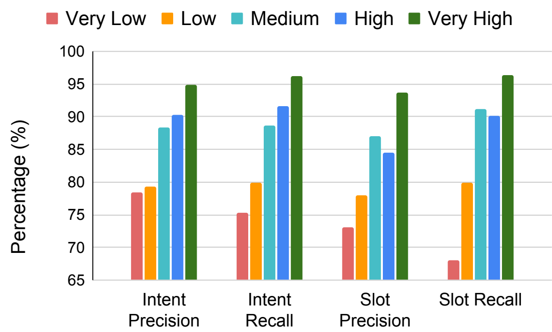

Figure 7: Precision and Recall of intents and slots w.r.t to frame frequency for semparse-nn of the RETRONLU.

Rare or uncommon frames are those example utterances whose ground truth semantic parses without slot value tokens appear infrequently in the training set. To analyze this, we divided the test set into five equal sizes i.e., *Very Low*, *Low*, *Medium*, *High*, and *Very High* sets, based on the frequency of

[12] The precision gap was small 1% (intents) and 4% (slots).

semantic frame structure. The experiment checks if performance improvement is mainly attributed to frequently repeating frames (frequent frames) or for rarely occurring frames (uncommon frames).

Analysis: Figure 6 shows that all models perform worse on rare frames. This is expected as the parametric model gets less data for training on these frames. Furthermore, many of the low-frequency frames are also complex utterances with more than one intent and have more slots too. Moreover, the nearest neighbour will be noisier for less frequent frames. This is evident from the lower values of precision (20% gap) and recall (25% gap) on the intent and slots for nearest neighbors in Figure 7.

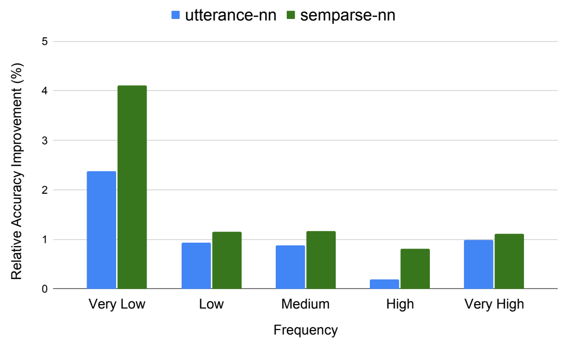

Figure 8: Relative performance improvement of RETRONLU w.r.t original baseline (without-nn) with varying frame frequency.

However, compared to original baseline (without-nn) the relative performance improvement on rare frames with retrieval augmented model is more substantial, as shown in Figure 8. For example, the relative improvement for *Very Low* frequency frames is 2.37% (utterance-nn) and 4.11% (semparse-nn) compared to just 1.01% (utterance-nn) and 1.11 % (semparse-nn)for the *Very High* Frequency frames. We suspect this is because of the model's ability to copy the required intent and slots from nearest neighbors if the parametric knowledge fails to generate it. This shows the retrieval augmented model is even more beneficial for the rare frames. As earlier, semparse-nn outperform utterance-nn.

Effect of the number of neighbors. To answer RQ5(a), we compare k = 1, 2, and 3 nearest neighbours for both utterance-nn and semparse-nn setups[13] The results are reported in Table 2.

[13]Extending beyond 3 neighbors was not useful for many reasons: (a) the BART 512 tokenization limit, (b) exponential rise in training time, and (c) only minimal performance gain.

#neighbors	k = 1	k = 2	k = 3
without-nn	84.43	84.43	84.43
utterance-nn	85.28	85.35	85.40
semparse-nn	**85.74**	**85.81**	**85.80**

Table 2: Performance with increasing nearest neighbors.

Metric	Average Precision		Average Recall	
Intent	**Farthest**	**Closest**	**Farthest**	**Closest**
Train	81.39	**84.84**	81.81	**85.04**
Valid	80.46	**87.59**	81.10	**87.93**
Test	79.09	**86.23**	79.35	**86.22**
Slot	**Farthest**	**Closest**	**Farthest**	**Closest**
Train	75.02	**80.05**	79.56	**83.19**
Valid	73.40	**82.38**	79.77	**85.81**
Test	74.59	**83.21**	79.51	**85.11**

Table 3: Intent-slots precision/recall for RETRONLU semparse-nn with closest/farthest neighbors.

Analysis: As shown in Table 2 the model performance only improves marginally with more nearest neighbors. We attribute this to the following two reasons (a) redundancy - many utterance examples can share the same frame, as evident from the high accuracy for frequent frame Figure 6., and (b) complexity - as k increases, the problem is getting harder for the model with longer inputs, more irrelevant and noisier inputs. To further verify the above reasons, we examine the semparse-nn retrieve nearest neighbors quality by comparing the intent and slot both Precision and Recall score for closest (k=1) and farthest (k=3) neighbor w.r.t to the gold semparse. From Table 3 it is evident that precision and recall for intents and slots decrease as we go down the ranked neighbors list. Adding more nearest neighbour would only be beneficial when added neighbour capture diverse and different semantic structure which is missing from earlier neighbor and essential for the correct semparse.

Effect of Retrieval Quality. To check if our RETRONLU model is robust to the noise in the retrieved examples (i.e. RQ5(b)), we analyse the effect of quality of retrieval by comparing semantic parsing accuracy of top neighbor augmented models on the test data with (a) the top neighbour with random neighbor from domain other than the example domain, and (b) random neighbor selected from the top 100 ranked nearest neighbors in the index. It should be noted that these 100 top rank nearest neighbour can have some redundant semparse-nn structure, only slot values might differ. Figure 9 shows the results of the experiments.

Analysis: From Figure 9 it is clear that quality of nearest neighbor affect the semantic parsing ac-

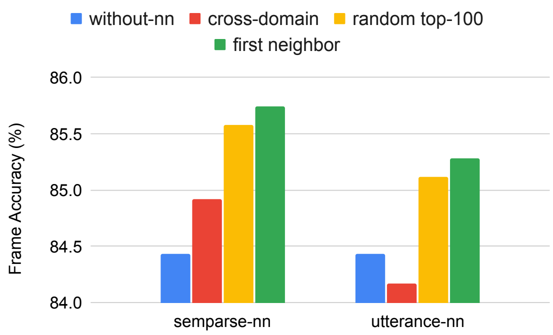

Figure 9: Performance of RetroNLU with varying nearest neighbor quality on test data.

curacy. We observe a 0.4% drop when random neighbor from top 100 nearest neighbors is chosen, instead of first neighbor, the small drop is because of redundancy in intent/slots structure between examples, only slots value could be major difference. However, the performance is still 0.9% to 1.0% better than the one without the nearest neighbor. We suspect this is because of the fact that the data has many utterances with similar semparse output. Upon deeper inspection we found that top-100 still includes many relevant frames, and therefore random examples from top-100 are often still relevant. Furthermore, there is also frame redundancy, many different utterance queries have similar semantic parse frames structure and only differ at the slot values. This is also evident from table 2 which shows adding more neighbors is not beneficial, because of frame redundancy. Surprisingly, we also observe that the model performance with random cross-domain neighbor is better than without-nn for semparse-nn by 0.5%. This shows that the model knows when to ignore the nearest neighbors and when to rely on the parametric model. Furthermore, it also indicates that underlying parametric model parameters is improved by retrieval augmented training for the semparse-nn.

For the utterance-nn the performance drops when testing on cross-domain nearest neighbor augmented example. Thus, underlying the utterance-nn model is more sensitive than semparse-nn to the nearest neighbor quality. In addition, we also conducted an experiment in which we added the best possible neighbor based on the gold parse frame structure. The trained model, though this approach was not robust and relies too heavily on coping frames from neighbors, resulting in poor generalization. Our technique, on the other hand, with embedding-based retrieval, is good at generalization and has enhances the underlying parametric model. Overall, we can conclude that the semparse-nn and utterance-nn model are both quite robust to nearest neighbors quality. We can also conclude that the semparse-nn model was able to capture richer information through additional similar inputs than without-nn. However, to obtain the best performance good quality neighbour is an essential.

4 Comparison with Related Work

Task-oriented Semantic Parsing. Sequence-to-sequence (*seq2seq*) models have recently achieved state of the art results in semantic parsing (Rongali et al., 2020; Gupta et al., 2018), and they also provide a flexible framework for incorporating session-based, complex hierarchical semantic parsing (Sun et al., 2019; Aghajanyan et al., 2020; Cheng et al., 2020; Mehri et al., 2020) and multilingual semantic parsing (Li et al., 2021; Louvan and Magnini, 2020). Architectures, such as T5 and BART (Raffel et al., 2020; Lewis et al., 2020b), with large pre-trained language models pushed the performance even further. Such models are quite capable of storing a lot of knowledge in their parameters (Roberts et al., 2020), and in this work we explore the benefits of additional non-parametric knowledge in a form of nearest neighbor retrieval for the task of semantic parsing. To improve low resource seq2seq parsers Chen et al. (2020) have proposed looking at meta learning methods such as reptile, and Ghoshal et al. (2021) have introduced new fine-tuning objectives. Our approach is focused on non-architecture changes to augment generation with retrieval and thus can be combined with either of these approaches.

Incorporating External Knowledge. An idea to help a model by providing an additional information, relevant to the task at hand is not new. This includes both implicit memory tables (Weston et al., 2014; Sukhbaatar et al., 2015), as well as incorporating this knowledge explicitly as an augmentation to the input. Explicit knowledge are incorporated in one of the following two ways (a) suitable model architecture change to incorporate dedicated extended memory space internally i.e. memory network (Bapna and Firat, 2019; Guu et al., 2020; Lewis et al., 2020a; Tran et al., 2020) or span pointer networks (Desai and Aly, 2021; Shrivastava et al., 2021), and (b) appending example specific extra knowledge externally with the input example directly without modifying model architecture (Papernot and McDaniel, 2018; Weston et al., 2018; Lewis et al., 2020c; Tran et al., 2020; Khandelwal et al., 2021; Fan et al., 2021; Chen et al., 2018; Wang et al., 2019; Neeraja et al., 2021) . Retrieval-augmented approaches have been improving language model pre-training as well (Guu et al., 2020; Lewis et al., 2020a; Tran et al., 2020). The idea here is to decouple memorizing factual knowledge and actual language modeling tasks, which can help mitigate hallucinations, and other common problems.

Multiple works like DkNN (Papernot and McDaniel, 2018), RAG (Lewis et al., 2020c), kNN-LM (Tran et al., 2020), kNN-MT (Khandelwal et al., 2021), and KIF-Transformer (Fan et al., 2021) show that external knowledge is useful for large pre-trained language models, and can help fine-tuning. DkNN shows that nearest neighbour augmented transformer-based neutral network is more robust and interpretable. RAG shows that one can append external knowledge to improve open domian, cloze-style question answering, and even fact verification task such as FEVER. kNN-LM shows that for cloze task for fact completion, one can combine nearest neighbour predictions with original prediction using appropriate weighting to improve model performance. However, these works mostly study knowledge dependent question answering task, while we are exploring a complex task of structural prediction of semantic frame structures for task-oriented dialog.

Very recently, Pasupat et al. (2021) share similar finding of exemplar augmentation and propose ControllAble Semantic Parser via Exemplar Retrieval (CASPER). In their work, the semantic parser gets relevant exemplars from a retrieval index, augments them with the query, and then generates an output parse using a generative seq2seq model. The exemplars serve as a control mechanism for the generic generative model: by modifying the retrieval index or the construction of the augmented query, one may alter the parser's behavior. Compare to them, our study focuses more on the influence of augmentation on the performance of the state-of-the-art Copy Transformer BART model for task-oriented semantic parsing. By design, the copy transformer effectively utilizes it's copy mechanism to get non-parametric information from augmented nearest neighbor semparse/utterances. Additionally, we conduct a detailed investigation of the influence of retrieval quality, utterance and semantic complexity, and the rarity of semantic frames. We anticipate that our findings will shed light on the potential advantages of retrieval enhancing parametric neural networks for the complex structural task of task-oriented semantic parsing.

5 Conclusion and Future Work

We show that task-oriented semantic parsing performance can be enhanced by augmenting neural model-stored parametric information with non-parametric external memory. On the TOPv2 dataset, we demonstrated that adding instances derived from a nearest neighbor index greatly improves the semantic parsing performance of a BART model with copy mechanism. Our RETRONLU model is able to achieve higher accuracy earlier with less training data (limited supervision setting), which allows maintaining a large index with annotated data, while using only a subset to train a model more efficiently. Lastly, we performed an analysis of performance improvements on different slices, and found RETRONLU to be more effective on rarer complex frames, compared to a traditional *seq2seq* model.

RETRONLU extensions, we focus on joint training of retrieval and parsing components. Having task specific utterances representation can benefit i.e. finding utterances with similar semantic parse. Exploring few/zero-shot performance could be interesting direction. Having an easily-updateable index enables you to amend annotations, add new ones, or remove existing ones, without affecting the model. It will be useful to study other approaches of sentence embedding, such as Reimers and Gurevych (2019). Finally, using cross-lingual representations such as mBART (Liu et al., 2020), could help multilingual semantic parsing.

Acknowledgements

We thank members of the Utah NLP group for their valuable insights and suggestions at various stages of the project; and reviewers their helpful comments. Additionally, we would like to express our gratitude to Xilun Chen, Asish Ghoshal, Arash Einolghozati, Shrey Desai, Anchit Gupta, Abhinav Arora, Sonal Gupta, Alexander Zotov, Ahmed Aly, and Luke Zettlemoyer of Meta (Formerly Facebook AI) for their insightful feedback and suggestions.

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)*, pages 5026–5035, Online. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. 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 1921–1931, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

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)*, pages 5090–5100, Online. Association for Computational Linguistics.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, Jason D. Williams, Hong Yu, Diarmuid Ó Séaghdha, and Anders Johannsen. 2020. Conversational semantic parsing for dialog state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117, Online. Association for Computational Linguistics.

Shrey Desai and Ahmed Aly. 2021. Diagnosing transformers in task-oriented semantic parsing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 57–62, 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.

Asish Ghoshal, Xilun Chen, Sonal Gupta, Luke Zettlemoyer, and Yashar Mehdad. 2021. Learning better structured representations using low-rank adaptive label smoothing. In *International Conference on Learning Representations*.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems*, volume 33, pages 18470–18481. Curran Associates, Inc.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick 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. 2020c.

Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

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*, pages 2950–2962, Online. Association for Computational Linguistics.

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 46–52, Marseille, France. European Language Resources association.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2799–2809, Online. Association for Computational Linguistics.

Nicolas Papernot and Patrick McDaniel. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*.

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.

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.

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*. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. *Don't Parse, Generate! A Sequence to Sequence Architecture for Task-Oriented Semantic Parsing*, page 2962–2968. Association for Computing Machinery, New York, NY, USA.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Akshat Shrivastava, Pierce Chuang, Arun Babu, Shrey Desai, Abhinav Arora, Alexander Zotov, and Ahmed Aly. 2021. Span pointer networks for non-autoregressive task-oriented semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1873–1886, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yibo Sun, Duyu Tang, Jingjing Xu, Nan Duan, Xiaocheng Feng, Bing Qin, Ting Liu, and Ming Zhou. 2019. Knowledge-aware conversational semantic parsing over web tables. In *Natural Language Processing and Chinese Computing*, pages 827–839, Cham. Springer International Publishing.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. volume 33, pages 7208–7215.

Jason Weston, Sumit Copra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

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.

A Qualitative Examples

Below, we provide some qualiative example from the TOPv2 datasets, in the first box, we shows the input utterance, nearest neighbour utterance i.e. nearest-nn utterance and it corresponding semantic parse i.e. semparse-nn (from the training set). In the second box we showed expected gold semantic parse (expected-sparse) and the semantic parse prediction with baseline i.e. without nearest neighbor (without-nn), and our models utterance only nearest neighbor, where the input is appending with nearest-nn utterance (utterance-nn) and semantic augmented model i.e. (semparse-nn), where the input is augmented with semantic parse of first nearest neighbor. We can clearly see from the examples that augmented nearest neighbor enhance models in identifying correct intent and slots.

example 1 input
utterance : message just lizzie and trent from my group if they have any updates yet ?

nearest-nn utterance : message kira and lena saying want to get drinks this week ?
nearest-nn semparse : [in:send_message message [sl:recipient kira] and [sl:recipient lena] saying [sl:content_exact want to get drinks this week]?]

example 1 output
expected-parse : [in:send_message [sl:recipient lizzie] [sl:recipient trent] [sl:content_exact they have any updates yet]]
without-nn : [in:get_message [sl:content_exact they have any updates yet] [sl:group lizzie] [sl:group trent]]
semparse-nn : [in:send_message [sl:recipient lizzie] [sl:recipient trent] [sl:content_exact they have any updates yet]]
utterance-nn : [in:send_message [sl:recipient lizzie] [sl:recipient trent] [sl:content_exact they have any updates yet]]

In example 1, the model misses the correct intent and corresponding slots completely, the correct intent is sending a message rather than receiving a message is correctly identified by both semparse-nn and utterance-nn.

example 2 input
utterance : no more country
nearest-nn utterance : no more music
nearest-nn semparse : [in:stop_music [sl:music_type music]]

example 2 output
expected-parse : [in:remove_from_playlist_music [sl:music_genre country]]
without-nn : [in:play_music [sl:music_genre country]]
semparse-nn : [in:remove_from_playlist_music [sl:music_genre country]]
utterance-nn : [in:remove_from_playlist_music [sl:music_genre country]]

In example 2, the baseline model without nearest neighbour did the exact opposite of intended task of removing music of genre country from the playlist. However, after augmenting nearest neighbor context the model quickly correct the expected intent and slot. It should also be noted the both the correct intent and slot (i.e. in:remove_from_playlist_music and sl:music_genre) are not present in the nearest-nn semparse but it do contain similar intent and slot (i.e. in:stop_music. and sl:music_type), which help retrieval augmented model in correct prediction. As earlier the model is able to predict correct even with utterance only augmentation too.

example 3 input
utterance : block all songs of mariah carey
nearest-nn utterance : delete mariah carey songs
nearest-nn semparse : [in:remove_from_playlist_music delete [sl:music_artist_name mariah carey] [sl:music_type songs]]

example 3 output
expected-parse : [in:remove_from_playlist_music [sl:music_artist_name mariah carey]]
without-nn : [in:unsupported_music [sl:music_type songs]]
semparse-nn : [in:remove_from_playlist_music [sl:music_type songs] [sl:music_artist_name mariah carey]]
utterance-nn : [in:remove_from_playlist_music [sl:music_type songs] [sl:music_artist_name mariah carey]]

In example 3 the model without nearest neighbor augmentation struggle to identify the intent from utterance text token “block" therefore prediction unsupported music as the intent and the music type as songs, however the model with augmented nearest neighbour example with “delete" intended slot

Percentage	10 %			20 %			30 %		
Domain	**w/o nn**	**uttr-nn**	**sem-nn**	**w/o nn**	**uttr-nn**	**sem-nn**	**w/o nn**	**uttr-nn**	**sem-nn**
Alarm	80.50	84.05	83.60	83.71	84.89	85.76	84.22	85.93	82.92
Event	68.56	78.33	79.38	75.01	80.85	82.32	77.64	81.91	82.92
Music	69.12	75.74	73.23	74.09	77.53	77.34	75.6	78.01	78.13
Timer	71.63	76.76	76.27	75.51	76.18	79.28	77.21	79.68	79.84
Navigation	74.30	73.86	76.44	77.89	79.40	79.96	80.11	81.79	81.61
Messaging	84.38	87.30	89.44	88.39	91.31	91.50	89.53	92.78	92.25

Table 4: Limited training setting results on various domain with original baseline (without-nn), RETRONLU model utterance-nn and semparse-nn, shown here as w/o nn, utter-nn and sem-nn respectively.

#neighbour's	one			two			three		
Domain	**w/o nn**	**uttr-nn**	**sem-nn**	**w/o nn**	**uttr-nn**	**sem-nn**	**w/o nn**	**uttr-nn**	**sem-nn**
Alarm	86.67	87.17	88.57	86.67	87.77	87.87	86.67	87.68	87.90
Event	83.83	85.03	84.77	83.83	84.92	85.26	83.83	85.26	85.34
Music	79.80	80.73	80.71	79.80	80.71	81.50	79.80	80.52	81.11
Timer	81.21	81.75	81.01	81.21	81.04	82.29	81.21	81.44	82.10
Messaging	93.50	94.52	94.65	93.50	94.92	95.05	93.50	94.88	94.92
Navigation	82.96	84.16	85.20	82.96	84.12	84.46	82.96	84.59	84.79

Table 5: Effect of number of nearest neighbours of RETRONLU performance across domains

correct identified both the intent and slots. Furthermore, using nearest neighbor augmentation, the model resolves the active passive voice confusion.

B Domain based Limited Training Setting

In Table 4 shows the performance of model for each domain on original baseline (without-nn), and RetroNLU model utterance-nn and semparse-nn with varying amount of supervised training data. Overall, semparse-nn outperform utterance-nn over most of the domains. Surprising, we also found that for few domain (with large number of samples) utterance-nn perform marginally better than semparse-nn, need to investigate exact reason for that. As expected both model utternace-nn and semparse-nn perform much better than original baseline which is without any nearest neighbour augmentation.

C Domain Specific Effect of Nearest Neighbours

In Table 5 we shows the performance of model for each domain on original baseline (without-nn), and RetroNLU model utterance-nn and semparse-nn with varying number of nearest neighbour augmented. We found the utternace-nn performance increases with increasing number of neighbours where semparse performance remain mostly constant after the first neighbour augmentation for many domains. We suspect this is due to the fact that the data contains a large number of utterances with identical semparse output.. There is also frame redundancy, since many unique utterance inquiries have comparable semantic parse frames structure with differences only on slot values.