Uncertainty and Traffic-Aware Active Learning for Semantic Parsing

Priyanka Sen
Amazon Alexa
sepriyan@amazon.com

Emine Yilmaz
Amazon Alexa
eminey@amazon.com

Abstract

Collecting training data for semantic parsing is a time-consuming and expensive task. As a result, there is growing interest in industry to reduce the number of annotations required to train a semantic parser, both to cut down on costs and to limit customer data handled by annotators. In this paper, we propose *uncertainty and traffic-aware active learning*, a novel active learning method that uses model confidence and utterance frequencies from customer traffic to select utterances for annotation. We show that our method significantly outperforms baselines on an internal customer dataset and the Facebook Task Oriented Parsing (TOP) dataset. On our internal dataset, our method achieves the same accuracy as random sampling with 2,000 fewer annotations.

1 Introduction

Semantic parsing is the task of mapping natural language to a machine-executable meaning representation. Supervised semantic parsing models are trained on corpora of natural language utterances with annotated meaning representations. Collecting these annotations is an expensive manual process, usually requiring expert annotators who are familiar with both the domain of utterances and the target meaning representation language (e.g. SQL).

Active learning is a method for collecting training data when annotating is difficult or budgets are limited (Settles, 2009). In active learning, an algorithm selects examples from an unlabeled set that are predicted to be more useful for the model if labeled. These examples are annotated and the model is retrained in an iterative process. The goal of an active learner is to reach higher performance faster than a random sampling baseline.

In this paper, we propose *uncertainty and traffic-aware active learning*, a simple yet effective method to improve a semantic parser. In our setup, we assume access to a set of initially annotated utterances and a large set of unlabeled utterances from customer traffic. We show that by using a combination of uncertainty and utterance frequency from traffic, we can achieve significantly higher performance than baselines on both an internal customer dataset and on the Facebook Task Oriented Parsing (TOP) dataset (Gupta et al., 2018).

2 Related Work

Active learning has been applied to various NLP tasks (Zhou et al., 2010; Li et al., 2012; Shen et al., 2017; Peshterliev et al., 2019; Chen et al., 2019). Duong et al. (2018) presented one of the first works on active learning for deep semantic parsing and found that selecting low-confidence examples outperformed random examples on two datasets but failed on a third. Koshorek et al. (2019) experimented with learning to actively-learn for semantic parsing, a method where the active learner is a learned model, but failed to see better performance than random sampling. Ni et al. (2020) proposed a framework where a weakly trained semantic parser was allowed to actively select examples for extra supervision. The authors found that selecting the least confident of the incorrect examples led to the best performance. Incorrect examples were identified by executing the predicted query and comparing the predicted answer with an expected answer. In this paper, we experiment with using uncertainty and utterance frequencies from customer traffic, a feature often found in industry logs.

3 Uncertainty and Traffic-Aware Active Learning

We propose *uncertainty and traffic-aware active learning* for semantic parsing. Our method is inspired by Mehrotra and Yilmaz (2015), who presented an active learning method for ranking al-

Proceedings of the First Workshop on Interactive and Executable Semantic Parsing, pages 12–17

https://doi.org/10.18653/v1/P17

gorithms which selects examples that are both informative to the model and representative of the dataset. The authors found that including a representativeness measure helped offset the tendency of informativeness measures to select outliers. In their paper, the authors measured informativeness as permutation probability based on a committee of ranking models, so a query where the most certain committee member had the least confidence was considered more informative. For representativeness, the authors used an LDA model to create a feature vector for each query. If a query's feature vector had higher cosine similarity to the average feature vector of all queries, the query was considered more representative.

In our method, we also use informativeness and representativeness, but we introduce new ways to measure both that can be applied to semantic parsing tasks. For each utterance u in a set of unlabeled utterances U, we calculate f(u), a sampling weight associated with u, as:

$$f(u) = \beta \frac{\phi(u)}{\sum_{u \in U} \phi(u)} + (1 - \beta) \frac{\psi(u)}{\sum_{u \in U} \psi(u)} \quad (1)$$

where $\phi(u)$ is the representativeness and $\psi(u)$ is the informativeness of u. We measure $\phi(u)$ as the utterance frequency, calculated as the number of times the utterance u appeared during a given time window of traffic. We measure $\psi(u)$ as 1 - our model's confidence on u. To calculate confidence, we use perplexity per word, which is the inverse probability of a model's output normalized by the number of words. We convert this perplexity into a confidence score by scaling it to a value between [0,1] using the function in Algorithm 1. The threshold is set to 0.9, which was fine-tuned based on the model's accuracy in production. In this function, confidence approaches 1 as perplexity approaches 0, confidence is 0.5 when perplexity is the threshold, and confidence approaches 0 as perplexity approaches infinity. While this scaled perplexity is not an exact measure of confidence, we found that it was effective in our experiments.

Both $\phi(u)$ and $\psi(u)$ are normalized by the sum of all values of $\phi(u)$ and $\psi(u)$. We use f(u) as a weight on each utterance when sampling. Utterances that maximize f(u) by having higher frequencies and lower confidences are more likely to be selected.

The β is a fine-tunable term that weighs the utterance frequency against the confidence. We manually fine-tuned β by training 9 models with different values ranging from 0.1 to 0.9 and compared performance in terms of exact-match accuracy. We found that a β of 0.4 performed the best on our internal dataset and a β of 0.5 performed the best on TOP, and so we use these β values in this paper.

Algorithm 1: Perplexity to confidence

```
p ← perplexity
if p > threshold: then
    return 1 / (2 + (100 * (p - threshold)));
else
    return 1 - 0.5 * (p / threshold);
end
```

3.1 Semantic Parsing Model

The semantic parsing model we use to evaluate our method is a reimplementation of the sequence-to-sequence model with pointer generator network proposed by Rongali et al. (2020), which achieved state-of-the-art performance on Facebook TOP (Gupta et al., 2018). We use a BERT-Base model (Devlin et al., 2019) as the encoder and a transformer based on Vaswani et al. (2017) as the decoder. The encoder converts a sequence of words into a sequence of embeddings. Then at each time step, the decoder outputs either a symbol from the output vocabulary or a pointer to an input token. A final softmax layer provides a probability distribution over all actions, and beam search maximizes the output sequence probability.

3.2 Compared Approaches

We compare our method to the following baselines.

RANDOM: Our random baseline randomly samples utterances for annotation.

TRAFFIC-AWARE: Our traffic-aware baseline uses utterance frequencies as a weight on each utterance, prioritizing utterances asked more often. In datasets containing duplicates, this is equivalent to random sampling.

CLUSTERING: In our clustering baseline (Kang et al., 2004; Ni et al., 2020), we compute a RoBERTa (Liu et al., 2019) embedding using sentence-transformers[1] for each utterance. We cluster the embeddings with

[1] https://github.com/UKPLab/sentence-transformers

	Internal	TOP
Train	10,000	500
Dev	2,000	4,032
Test	5,000	8,241
Unlabeled	100,000	13,680
Src Vocab	30,160	11,873
Tgt Vocab	5,400	116

Table 1: Details of the datasets. *Train* is the starting training set in our experiments. *Unlabeled* is the set from which additional training examples are sampled.

Internal	what is the capital of france, `is_the_capital_of(@ptr5)`

TOP	Any accidents along Culver,
	`[IN:GET_INFO_TRAFFIC`
	`@ptr0 @ptr1 @ptr2`
	`[SL:LOCATION @ptr3]]`

Table 2: Examples from the datasets. `@ptrs` are pointers to a source token. In the first example `@ptr5` refers to the 5th token in the source, "france".

k-means and set the number of clusters to the round's budget (i.e. if our budget is 500 utterances, we create 500 clusters). Then we randomly sample 1 example per cluster.

LEAST CONFIDENCE: Our least confidence baseline (Lewis and Catlett, 1994; Culotta and McCallum, 2005) selects utterances with the lowest model confidence.

MARGIN OF CONFIDENCE: Our margin of confidence baseline (Settles and Craven, 2008) calculates the difference in confidence between the top two predictions in an n-best list. Large differences between the top two predictions indicate there is a clear top prediction, while small differences indicate greater model uncertainty. We select the examples with the smallest difference in confidence.

UNCERTAINTY-AWARE: A less deterministic version of Least Confidence. We use 1 - model confidence as a weight on each utterance, prioritizing utterances with low confidence.

UNCERTAINTY + CORRECTNESS: Our uncertainty + correctness baseline (Ni et al., 2020) selects the most uncertain of the predictions that are incorrect. In practice, there are several ways to identify an incorrect prediction, such as checking if 1) a query fails to execute, 2) a query executes but fails to answer, or 3) a query executes but does not return the expected answer. In our experimental setup, we use a more favorable setting by checking the prediction against the expected representation.

4 Datasets

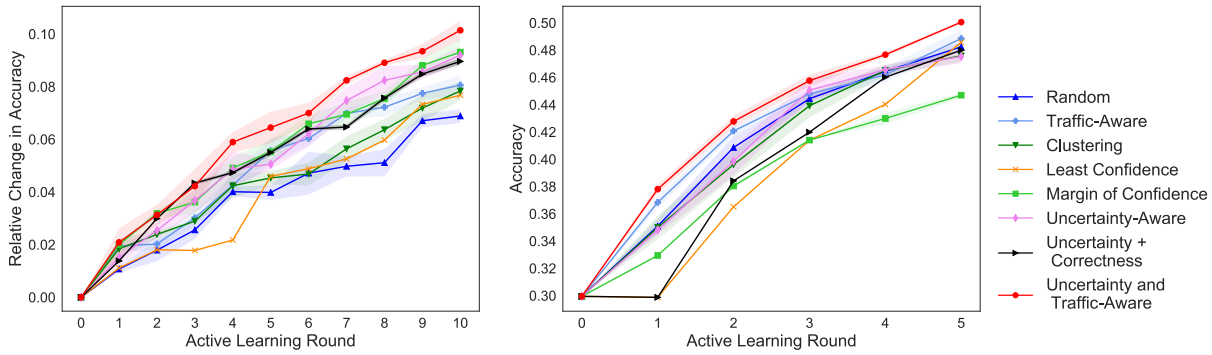
We run experiments on both an internal customer dataset and the Facebook Task Oriented Parsing (TOP) dataset (Gupta et al., 2018). Details and examples are shown in Tables 1 and 2.

Our internal dataset contains open-domain factual questions asked by customers to a commercial voice assistant. The utterances are anonymized and labeled with a meaning representation by an internal high-precision rule-based system. We also calculate a count for each utterance based on how often the utterance was asked in a given period of time. This dataset contains only unique utterances, which prevents selecting the same utterance multiple times for annotation.

To our knowledge, there is no public semantic parsing dataset with question frequencies, and so we use a modified version of TOP. TOP is a semantic parsing dataset of 45k crowdsourced queries about navigation and public events. These queries are manually labeled with a meaning representation. In order to create a measure of representativeness, we assume that utterances with an exact-matched meaning representation are semantically similar. Utterances with meaning representations that appear more often are considered more representative. We keep one utterance per exact-matched meaning representation, and use the counts as a measure of how popular this type of question is among users. This is done for experimental purposes. In a real setting without the labels, we could use alternate measures of semantic similarity to identify more popular questions.

5 Experiments

For controlled experimentation, we simulate active learning by treating a subset of our data as unlabeled. When an unlabeled example is selected, we reveal the label and add it to the training set. All our experiments are run on an Nvidia Tesla v100 16GB GPU and the results are reported as exact match accuracy.



(a) Internal

(b) TOP

Figure 1: Results of the experiments. Scores are calculated as exact-match accuracy. We only report relative change in accuracy for the internal dataset. The shaded regions represent the standard error for each point.

5.1 Internal Dataset

For our internal dataset, we start with a base training set of 10,000 utterances and set an annotation budget of 5,000 utterances. In each round, we sample 500 utterances from the unlabeled set, append them with their labels to the training set, and fully retrain the model. We repeat this for 10 rounds and report results as an average over 5 runs.

The results are shown in terms of relative change in exact-match accuracy in Figure 1a. Our method initially has similar performance to uncertainty-based baselines, but after Round 4, our method outperforms all the baselines. Table 3 has results of paired t-tests comparing our method to each baseline. All the p-values are <0.05, showing statistical significance. In particular, our method outperforms random sampling. The examples picked by the first 6 rounds of uncertainty and traffic-aware sampling (accuracy $\Delta 7.0\%$ at round 6) are as valuable as the examples picked by all 10 rounds of random sampling (accuracy $\Delta 6.9\%$ at round 10), saving on the cost of 2,000 annotations.

To better understand these results, we inspected examples selected by each method. We found that although the traffic-aware method picked popular utterances, annotating many similar questions had limited gains over time. On the other hand, uncertainty-based approaches picked more diverse examples, but since customer datasets can be noisy, they were prone to picking outliers that were not as useful to the model when annotated. By combining frequency with uncertainty, our method was able to prioritize popular but under-represented examples, which were both interesting for customers and interesting for the model, and this gave us the best performance.

Baseline	Internal	TOP
Random	$p<.001$	$p=.01$
Traffic-Aware	$p<.001$	$p=.008$
Clustering	$p<.001$	$p=.01$
Least Confidence	$p<.001$	$p=.02$
Margin of Confidence	$p=.002$	$p=.004$
Uncertainty-Aware	$p<.001$	$p=.02$
Uncertainty + Correctness	$p=.001$	$p=.03$

Table 3: Results of paired t-tests comparing our method to each baseline. $p<.05$ is considered significant

5.2 TOP

We next ran experiments on TOP. Given that TOP is a smaller and simpler dataset (e.g. target vocab of 116 vs. 5,400), we start with a smaller base training set of 500 examples and set an annotation budget of 500 examples. In each round, we sample 100 examples from the unlabeled set, append them with their labels to the training set, and fully retrain the model. We see the effect of our method as early as Round 1, so we stop after 5 rounds and report results as an average over 5 runs.

The results are shown as exact-match accuracy in Figure 1b and the p-values from paired t-tests are in Table 3. These results again show that our method significantly outperforms the baselines. Even though the traffic weights in TOP are not from customer traffic, traffic-aware sampling performs almost as well as our method. This suggests that MRL frequency is a helpful measure for this test set. We also observe that some of our uncertainty-based baselines perform worse than random sampling, in contrast to our results on the internal dataset. We hypothesize this could be because uncertainty is a

less useful signal from models built with smaller training sets (TOP: 500-1,000 training examples vs. Internal: 10,000-15,000 training examples) or because low confidence examples were less useful for TOP's test set. Uncertainty still provides some advantage, however, as the combination with MRL frequency leads to the best performance.

6 Conclusion

In this work, we present *uncertainty and traffic-aware active learning*, a method that uses model confidence and traffic frequency to improve a semantic parsing model. We show that our method significantly outperforms baselines on both an internal dataset and TOP. Our method achieves the same precision as random sampling with 2,000 fewer annotations on our internal dataset. Based on our results, we present our method as a way to improve semantic parsers while reducing annotation costs and limiting customer data shown to annotators.

References

Xi C. Chen, Adithya Sagar, Justine T. Kao, Tony Y. Li, Christopher Klein, Stephen Pulman, Ashish Garg, and Jason D. Williams. 2019. Active Learning for Domain Classification in a Commercial Spoken Personal Assistant. In *Proc. Interspeech 2019*, pages 1478–1482.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2018. Active learning for deep semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 43–48, Melbourne, Australia. Association for Computational Linguistics.

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.

Jaeho Kang, Kwang Ryel Ryu, and Hyuk-Chul Kwon. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 384–388. Springer.

Omri Koshorek, Gabriel Stanovsky, Yichu Zhou, Vivek Srikumar, and Jonathan Berant. 2019. On the limits of learning to actively learn semantic representations. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 452–462, Hong Kong, China. Association for Computational Linguistics.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pages 148–156. Elsevier.

Shoushan Li, Shengfeng Ju, Guodong Zhou, and Xiaojun Li. 2012. Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 139–148, Jeju Island, Korea. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rishabh Mehrotra and Emine Yilmaz. 2015. Representative & informative query selection for learning to rank using submodular functions. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 545–554, New York, NY, USA. Association for Computing Machinery.

Ansong Ni, Pengcheng Yin, and Graham Neubig. 2020. Merging weak and active supervision for semantic parsing. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, New York, USA.

Stanislav Peshterliev, John Kearney, Abhyuday Jagannatha, Imre Kiss, and Spyros Matsoukas. 2019. Active learning for new domains in natural language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 90–96, Minneapolis, Minnesota. 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. In *Proceedings of The Web Conference 2020*, pages 2962–2968.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Shusen Zhou, Qingcai Chen, and Xiaolong Wang. 2010. Active deep networks for semi-supervised sentiment classification. In *COLING 2010: Posters*, pages 1515–1523, Beijing, China. COLING 2010 Organizing Committee.