Ambiguous Learning from Retrieval: Towards Zero-shot Semantic Parsing

Shan Wu[1,3,*], Chunlei Xin[1,3,*], Hongyu Lin[1,†], Xianpei Han[1,2], Cao Liu[4], Jiansong Chen[4], Fan Yang[4], Guanglu Wan[4], Le Sun[1,2,†]

[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China [4]Meituan-Dianping Group
{wushan2018,chunlei2021,hongyu2016,xianpei,sunle}@iscas.ac.cn,
{liucao,chenjiansong,yangfan79,wanguanglu}@meituan.com

Abstract

Current neural semantic parsers mostly take supervised approaches, which require a considerable amount of expensive training data. As a result, minimizing supervision requirements has been one of the key challenges in semantic parsing. In this paper, we propose a Retrieval as Ambiguous Supervision framework, which can effectively collect high-coverage ambiguous supervisions (i.e., the parse candidates of an utterance) via a pre-trained language models-based retrieval system. Then, by assuming candidates will contain the correct ones, the zero-shot task can be converted into an ambiguously supervised task. To improve the precision and coverage of such ambiguous supervision, we propose a confidence-driven self-training algorithm, in which a semantic parser is learned and exploited to disambiguate candidates iteratively. Experimental results show that our approach significantly outperforms the state-of-the-art zero-shot semantic parsing methods.

1 Introduction

Semantic parsing aims to map natural language sentences into computer-understandable meaning representations(MRs), which has attracted substantial attention for many years (Wong and Mooney, 2007; Kate et al., 2005; Lu et al., 2008; Dong and Lapata, 2016). Nowadays, neural network methods have become the mainstream for semantic parsing. Since neural semantic parsers are limited to the patterns observed in the training data, a large number of annotated data is required. However, annotating utterances with detailed, correct meaning representations is a difficult and time-consuming task, which relies on expert knowledge about MRs.

Recent studies in semantic parsing try to employ pre-trained language models (PLMs) to alleviate the problem of data insufficiency. Shin et al. (2021); Wu et al. (2021); Schucher et al. (2022) reformulate semantic parsing as constrained paraphrasing generation, where paraphrasing generation is modeled by PLMs. To eliminate the need for human-annotated data, Xu et al. (2020) employ PLMs to paraphrase repeatedly and obtain millions of data. However, these methods still rely on lots of detailed annotated data or heavy data synthesis.

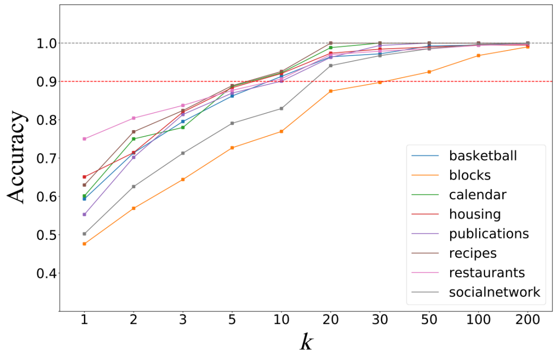

Figure 1: The top-k accuracies of the retrieved MRs by PLMs-based retriever on the eight domains in OVERNIGHT. We can see that the retrieved results have high top-k accuracy but low precision.

In this paper, we propose a Retrieval as Ambiguous Supervision (RaAS) framework for zero-shot semantic parsing, which is simple and effective. In the RaAS framework, we make full use of a PLMs-based retriever to return high-coverage candidates, and then convert zero-shot semantic parsing into ambiguously supervised semantic parsing[1]. As previous work found, sentence similarity and PLMs can provide effective candidates: Herzig and Berant (2019) use sentence similarity scores and Belyy et al. (2022) use PLMs to provide candidates for manual annotation, and PLMs-based paraphrasing models can provide parsing results with consider-

*Equally Contribution.
†Corresponding Authors.

[1]In ambiguous supervision (Kate and Mooney, 2007; Kim and Mooney, 2010), where each sentence is annotated with multiple potential meaning representations and the correct ones are within them. Strictly speaking, our setting is approximate ambiguous supervision or noisy ambiguous supervision.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14081–14094

able top-20 accuracy (Wu et al., 2021). Thus, we propose an effective PLMs-based retrieval system to retrieve MRs from the collected MRs datastore, and select the top-k MRs as ambiguous supervision signals, in which we suppose there is at least one true meaning representation. Then, we employ a self-training protocol that exploits the sequences modeling ability of semantic parsers to improve the coverage and precision of candidates. In our approach, semantic parsers are learned and exploited to supplement candidates and disambiguate the MRs iteratively.

Without any supervision, our PLMs-based retrieval system can provide discriminative supervision signals. In our retrieval system, the MRs datastore is built by sampling MRs under a limited depth and preserving the valid ones. Following previous work (Berant and Liang, 2014; Cao et al., 2020), we canonicalize the MRs for scoring. The sentence similarity scores between the query and canonical utterances are calculated by PLMs to retrieve MR candidates. As shown in Fig 1, the retrieval results of PLMs have high top-k accuracy. In all domains of OVERNIGHT, the average top-20 accuracy can reach 95.3% but the average top-1 accuracy is only 59.5%. We assume that the retrieval results can provide sufficient ambiguous supervision, of which the precision and coverage can be further improved by SEQ2SEQ models.

To further improve the precision and coverage of the above ambiguous supervision, we propose a confidence-driven self-training algorithm. Our learning method iterates between two stages: 1) Train the semantic parser from the high confidence instances; 2) Expand candidate sets and update the confidence weights of candidates based on the current parser.

In summary, our main contributions are:

- We propose the Retrieval as Ambiguous Supervision framework, which can exploit the prior knowledge of PLMs and the sequences modeling ability of semantic parsers simultaneously.
- We design a confidence-driven self-training algorithm on retrieval, which can improve the precision and coverage of ambiguous supervision.
- Experiments on three standard datasets show that our approach significantly outperforms previous zero-shot semantic parsing methods.

2 Retrieval as Ambiguous Supervision Framework

We propose Retrieval as Ambiguous Supervision framework, which treats the retrieval results as ambiguous supervision signals (Fig. 2). First, for each sentence, we use a pre-trained model to provide reliable meaning representation candidates, in which we assume that at least one is correct. So the zero-shot semantic parsing is converted into an ambiguous supervision task. Then we propose a confidence-driven self-training algorithm, in which high-confidence instances from the candidates are used to train the semantic parser and in turn the semantic parser is exploited to supplement and disambiguate the candidates. This process is iterative.

2.1 PLMs-based MRs Retrieval System

In order to make better use of the PLMs to retrieve semantic parsing candidates, we first use the production rules of meaning representations and the constraints of knowledge base to build the retrieval datastore D. Then, given a query sentence x, the pre-trained language models are used to calculate the retrieval score for each MR y in D. The top-k retrieval results form the candidate sets U_x, which are viewed as ambiguous supervision signals.

2.1.1 MRs Collecting

For each domain, we use the context free grammar (CFG) of the corresponding semantic formalism. We randomly expand the production rules of CFG to sample a large number of meaning representations Y'. To make full use of the knowledge constraints of the knowledge base, we only preserve the executable meaning representations Y.

Following previous work (Jia and Liang, 2016; Xu et al., 2020), through synchronous grammar, we also produce canonical utterances, which are the pseudo-language representations of MRs. Finally, we collect accessible meaning representation and canonical utterance pairs $\langle y, z\rangle$ to build retrieval datastore $D = \{\langle y_1, z_1\rangle, \langle y_2, z_2\rangle, ..., \langle y_n, z_n\rangle\}$.

2.1.2 PLMs-based Retriever

Following previous studies (Su and Yan, 2017; Cao et al., 2020; Wu et al., 2021), we first use canonical utterances to calculate retrieval scores. Canonical utterances can be viewed as sub-language representations of MRs. There is a one-to-one mapping between them. Formally, each MR y can be mapped to its cannonical utterance z by synchronous gr-

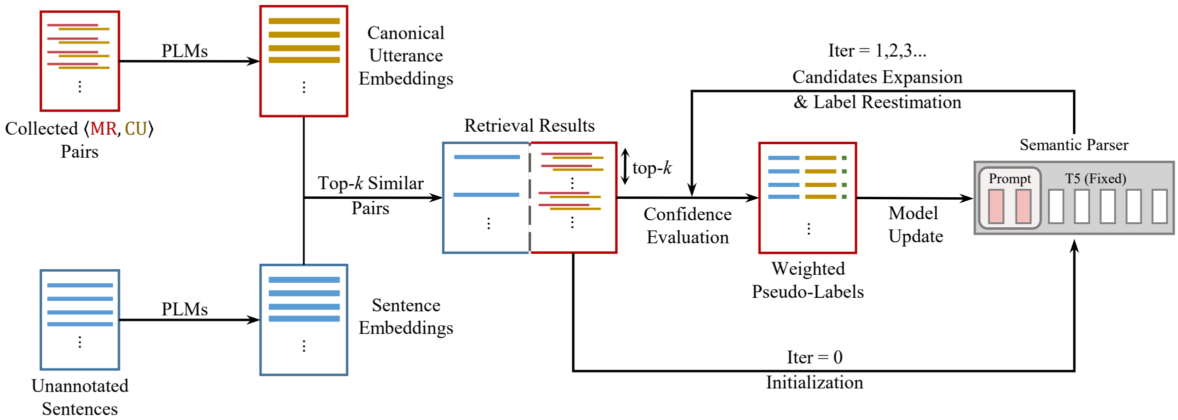

Figure 2: Our Retrieval as Ambiguous Supervision framework. The blue, red and yellow lines represent sentences, meaning representations and canonical utterances respectively. Green dots indicate the weights of the paired sentence and canonical utterance instances. The T5 model is fixed, only the soft prompt (pink parts) is fine-tuned.

Query Utterance:

Which player had the same amount of assist as Kobe Bryant

Retrieval:

#1 × Score: 0.9390689

Number of assist of player Kobe

property (property (kobe, reverse (player)), num_assists)

#2 √ Score: 0.9367738

Player whose number of assist is number of assist of player Kobe

property ((λ s (filter (s , num_assists = property (property (kobe , reverse (player)), num_assists)))), player)

#3 × Score: 0.9365325

Number of assist of player Kobe whose season is 2004

property (filter (property (kobe, reverse (player)), season = 2004), num_assists)

Figure 3: An example of the retrieval results from our PLMs-based retriever.

mmar. We use z to compute the retrieval score $r_{x,y}$.

Given a query sentence x, we can calculate the cos similarity of x and each canonical utterance z in D by $\cos(\mathbf{h}(x), \mathbf{h}(z))$ with the PLM encoder $\mathbf{h}$. The encoder has been pre-trained on large-scale public datasets in advance and has not touched any canonical utterances. We normalize the cos similarities to calculate the scores:

$$score_{\mathbf{h}}(x, z) = \frac{e^{\cos(\mathbf{h}(x),\mathbf{h}(z))/\tau}}{\sum_{\langle y',z'\rangle \in D} e^{\cos(\mathbf{h}(x),\mathbf{h}(z'))/\tau}} \quad (1)$$

, in which τ is the temperature parameter. The initial confidence scores are obtained from the similarities: $r_{x,y} = score_{\mathbf{h}}(x, z)$. We keep the top-$k$ retrieval results $U_x = [\langle y_1, z_1\rangle, \langle y_2, z_2\rangle, ..., \langle y_k, z_k\rangle]$ and their corresponding scores for later ambiguous learning. In our practice, k is set to 20.

Although the retrieval system can provide discriminative supervision signals, the coverage and precision of MR candidates should be further refined. As shown in the example of Fig 3, the retrieval system pays more attention to the relevance and confuses the highly relevant utterances. In this example, the related words 'player', 'amount', 'assist' and 'Kobe' all appear in the first and second candidates, but the meanings of the correct MR #2 and #1 are very different. This demonstrates that the retrieval model does not have enough understanding of their accurate semantics. However, it still provides a good initialization of candidates and confidence scores, which can be further refined by more accurate SEQ2SEQ modeling.

2.2 Self-training on Retrieval

As mentioned above, after obtaining the ambiguous supervision signals U_x for each given input x and their corresponding initial confidence scores r, we propose a confidence-driven self-training protocol to improve the coverage and precision of candidates with SEQ2SEQ modeling. Our self-training algorithm operates in an EM-like manner, iterating between two stages: 1) Train a semantic parser from the candidates based on their confidence scores. 2) Exploit the current parser to expand the candidates and re-estimate their confidence scores;

In our self-training protocol, the Seq2Seq parser with semantic mapping ability is fed with reliable guidance from high-confidence instances, to denoise the supervision of relevant instances iteratively. As shown in Fig 4, after self-training iterations, the parser learns that 'Which player' maps to 'player' rather than 'number' and re-estimates the confidence scores to raise the ranking of the correct MR consequently. Thus the quality of supervision signals can be improved in such iterative

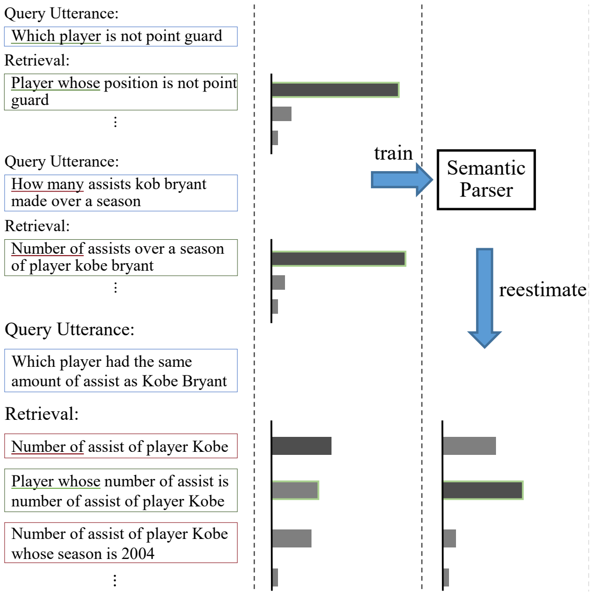

Figure 4: As mentioned above, the retrieval system pays more attention to the relevance, which confuses highly relevant utterances. After self-training iterations, the parser trained on high-confidence instances learns that 'Which player' queries 'player' rather than 'number' and improves the ranking of the correct answer.

re-estimation, which continually produces better parsers.

2.2.1 Prompt-based Semantic Parsers

As shown in previous work (Lester et al., 2021; Schucher et al., 2022), the prompt tuning is suitable for solving the overfitting problem in low resource settings. Following them, we use T5(Raffel et al., 2020) as the base model, and set the prompt length to 150.

Given a tokenized utterance $x = [x_1, x_2, ..., x_n]$, T5 encodes x into $E_x \in \mathcal{R}^{n\times e}$, where e is the dimension of the embedding space. The soft prompt is represented as a parameter $\theta_p = [P_1; P_2; ...; P_v] \in \mathcal{R}^{v\times e}$, in which v is the length of the prompt. The soft prompt is prepended to the input embeddings as $[\theta_p; E_x]$, which is provided to the language model. During prompt tuning, we only optimize θ_p, and fix the model parameters and the pre-trained vocabulary embeddings of T5.

Before self-training iterations (in Iter0), we use the top-1 of the retrieval results U_x as supervision signals to initialize the semantic parsing model.

2.2.2 Candidate Expansion and Confidence Re-estimation

In order to improve the precision and coverage of retrieval results, we add the top-m parsing results to the candidate set and disambiguate meaning representation annotations in a moving-average style after each model update.

Candidate Expansion As mentioned above, the ambiguous supervision can only be retrieved from the collected data. To make up for the generation label space, the m-best beam search results of the current semantic parser in t-th iteration $Y_x^t = [\langle y_1, z_1\rangle, \langle y_2, z_2\rangle, ..., \langle y_m, z_m\rangle]$ are employed to update the candidate set: $U_x^t = U_x \cup Y_x^t \ (t \geq 1)$.

Confidence Re-estimation To improve the supervision precision, and especially to resolve the problem that the retrieval system focuses more on relevance than on precise semantics, we use the generation model to refine and re-estimate the confidence $s_{x,y}^t$ of MR labels.

We first use a pre-trained paraphrase generation model g to refine the confidence scores:

$$s_{x,y}^0 = r_{x,y} + \frac{p_{\mathbf{g}}(x|z)}{\sum_{\langle y',z'\rangle \in U_x} p_{\mathbf{g}}(x|z')} \quad (2)$$

After each model update, we use the new parser $p(y|x)$ to re-evaluate the confidence scores of the meaning representation candidates in a moving-average style:

$$s_{x,y}^t = (1-\alpha)\frac{p(y|x)}{\sum_{y' \in U_x^t} p(y'|x)} + \alpha s_{x,y}^{t-1} \quad (3)$$

For the meaning representations newly added to the candidate set, we re-estimate their confidence scores as: $s_{x,y}^t = (1-\alpha^t)\frac{p(y|x)}{\sum_{y' \in U_x^t} p(y'|x)} + \alpha^t(r_{x,y} + \frac{p_{\mathbf{g}}(x|z)}{\sum_{\langle y',z'\rangle \in U_x^t} p_{\mathbf{g}}(x|z')})$.

Finally, we get the normalized confidence scores $S_t(y|x)$ as:

$$S_t(y|x) = \frac{s_{x,y}^t}{\sum_{y' \in U_x^t} s_{x,y}^t} \quad (4)$$

2.2.3 Self-training Update on Retrieval

Our learning framework operates in an EM-like manner, iterating between two stages: 1) Add candidates and update the confidence weights of the candidates based on current model parameters; 2) Train the parser from the soft pseudo instances. In the iterations, candidate samples are weighted to train the parser.

We use the continuous self-training method proposed by Zou et al. (2019). First, according to the normalized confidence $S_t(y|x)$, we resolve the soft pseudo-labels as:

$$\hat{y}_x^t = \underset{\hat{y}_x}{\text{argmin}} - \sum_{y \in U_x^t} \hat{y}_{x,y} \log S_t(y|x) + \beta r(\hat{y}_x) \tag{5}$$

, in which $\hat{y}_x^t \in \Delta^{|U_x^t|-1}$. We use a negative entropy label regularizer $r(\hat{y}_x) = \sum_{y \in U_x^t} \hat{y}_{x,y} \log \hat{y}_{x,y}$. The distribution of labels can be solved as:

$$\hat{y}_{x,y}^t = \frac{S_t(y|x)^{1/\beta}}{\sum_{y' \in U_x^t} S_t(y'|x)^{1/\beta}} \tag{6}$$

According to the weights of the candidate annotations, we train the parser by the following loss function:

$$\mathcal{J}(x, U_x^t) = - \sum_{y \in U_x^t} \hat{y}_{x,y}^t \log p(y|x; \theta_p) \tag{7}$$

2.2.4 Inference

When inferring, we follow the same way as confidence re-estimation. Given a query x, the candidate set consists of retrieval results and beam search results: $U = U_x \cup Y_x$. Then, we use the similar confidence re-estimation algorithm as in self-training: $score(x, y) = \frac{p(y|x)}{\sum_{y' \in U_x} p(y'|x)} + s_{x,y}^0$ to rerank candidates.

Following previous studies (Wu et al., 2021; Shin et al., 2021), we employ constrained decoding and generate canonical representations over meaning representations.

3 Experiments

Datasets We conduct experiments on three datasets: OVERNIGHT(λ-DCS), GEOGRANNO, and GEO(FunQL), which use different meaning representations and are on different domains. Note that we do not use any MR annotations in training set.

OVERNIGHT This is a dataset across eight domains, which contains natural language paraphrases paired with lambda DCS logical forms. We use the same train/test splits as Wang et al. (2015).

GEOGRANNO This is a semantic parsing benchmark about U.S. geography (Herzig and Berant, 2019), in which lambda DCS logical forms paired with canonical utterances are produced from SCFG. Instead of paraphrasing sentences, crowd workers are required to select the correct canonical utterance from candidate list. We follow the split (train/valid/test 487/59/278) in original paper.

GEO(FunQL) This is another version of GEO (Zelle and Mooney, 1996) using the variable-free semantic representation FunQL (Kate et al., 2005). We extend the FunQL grammar to SCFG for this dataset. Different from the previous datasets, the construction method of this dataset does not dependent on paraphrasing, which can better verify the effectiveness of our methods. We follow the standard 600/280 train/test splits.

Pretrained Language Models We use the pretrained sentences similarity model MPNet[2] (Song et al., 2020) as the retrieve model. The paraphrase generation model is the PEGASUS model (Zhang et al., 2020) fine-tuned for paraphrasing[3]. The PLMs have been trained on the public paraphrase datasets, which have not touched any canonical utterances. In our experiments, they are fixed and only used for retrieval and reranking.

System Settings We train all our models with 3 self-training iterations. In each iteration, the neural semantic parser is trained 1000 epochs, with the initial prompt learning rate of 0.1. We use Adam algorithm to update parameters, with batch size as 80 ~250. The temperature parameter τ is set to 0.1. We initialize soft prompt parameters by uniformly sampling within [-0.1, 0.1]. The beam size m during decoding and candidates expanding is 8. The hyper-parameters α is set to 0.5, β is set to 0.1.

Datastore Collecting We use synchronous context free grammars (SCFGs) to generate $\langle$MR, CU$\rangle$ pairs in each dataset. We generate roughly 800K, 250K, 20K pairs in OVERNIGHT, GEOGRANNO, GEO(FunQL) respectively. We only preserve the valid ones (are executable or meet type checking), and remove the redundant MRs. We collect roughly 10K, 20K, 3K valid pairs for our datastore in these datasets.

Few-shot Settings Following the previous few-shot settings in OVERNIGHT (Shin et al., 2021; Schucher et al., 2022), we randomly subsample 200 training examples for each domain as supervise data, and 20% of the remaining data is used for validation. All other data in training sets are treated as unannotated data, whose ambiguous supervision signals also come from the retrieval results.

[2]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[3]https://huggingface.co/tuner007/pegasus_paraphrase

	Bas.	Blo.	Cal.	Hou.	Pub.	Rec.	Res.	Soc.	Avg.
Supervised									
RECOMBINATION (Jia and Liang, 2016)	85.2	58.1	78.0	71.4	76.4	79.6	76.2	81.4	75.8
CROSSDOMAIN (Su and Yan, 2017)	86.2	60.2	79.8	71.4	78.9	84.7	81.6	82.9	78.2
SEQ2ACTION (Chen et al., 2018)	88.2	61.4	81.5	74.1	80.7	82.9	80.7	82.1	79.0
DUAL (Cao et al., 2019)	87.5	63.7	79.8	73.0	81.4	81.5	81.6	83.0	78.9
TWO-STAGE (Cao et al., 2020)	87.2	65.7	80.4	75.7	80.1	86.1	82.8	82.7	80.1
SSD (Wu et al., 2021)	86.2	64.9	81.7	72.7	82.3	81.7	81.5	82.7	79.2
Few-shot									
GPT-3 (Shin et al., 2021)	85.9	63.4	79.2	74.1	77.6	79.2	84.0	68.7	76.5
T5-base (Schucher et al., 2022)	78.6	45.2	68.2	63.6	67.5	70.5	73.3	61.4	66.0
T5-large (Schucher et al., 2022)	81.9	52.5	76.8	71.2	74.4	78.9	76.9	65.5	72.3
T5-xl (Schucher et al., 2022)	83.9	54.4	77.7	72.9	77.0	79.1	78.9	70.2	74.3
RaAS (w/o Self-Training)	78.0	51.9	70.2	68.8	67.1	71.3	78.9	61.8	68.5
RaAS (Full Model)	78.5	57.1	72.0	76.7	74.5	72.7	86.1	63.0	72.6
Zero-shot									
Cross-domain Zero Shot (Su and Yan, 2017)	-	28.3	53.6	52.4	55.3	60.2	61.7	-	-
GENOVERNIGHT (Wang et al., 2015)	15.6	27.7	17.3	45.9	46.7	26.3	61.3	9.7	31.3
WMDSAMPLES (Cao et al., 2020)	31.9	29.0	36.1	47.9	34.2	41.0	53.8	35.8	38.7
TWO-STAGE (Cao et al., 2020)	64.7	53.4	58.3	59.3	60.3	68.1	73.2	48.4	60.7
AUTOQA (Xu et al., 2020)	73.9	54.9	**72.6**	70.9	**74.5**	68.1	78.6	**61.5**	69.4
SSD (Wu et al., 2021)	71.3	**58.8**	60.6	62.2	58.8	65.4	71.1	49.1	62.2
RaAS (Retriever)	59.3	47.6	60.1	65.1	55.3	63.0	75.0	52.8	59.8
RaAS (w/o Self-Training)	61.1	51.6	64.3	66.7	62.1	64.8	75.9	52.7	62.4
RaAS (Full Model)	**78.0**	55.6	71.4	**76.7**	73.9	**71.3**	**85.5**	58.6	**71.4**

Table 1: Overall results on OVERNIGHT.

	GEO GRANNO	GEO (FunQL)
Supervised		
DEPHT (Jie and Lu, 2018)	-	89.3
COPYNET (Herzig and Berant, 2019)	72.0	-
One-stage (Cao et al., 2020)	71.9	-
Two-stage (Cao et al., 2020)	71.6	-
SEQ2SEQ (Guo et al., 2020)	-	87.1
SSD (Wu et al., 2021)	72.9	88.3
Unsupervised		
SYNTH-SEQ2SEQ (Wu et al., 2021)	32.7	36.1
WMDSAMPLES (Cao et al., 2020)	35.3	-
Two-stage (Cao et al., 2020)	63.7	-
SSD (Wu et al., 2021)	58.5	63.2
SSD-SAMPLES (Wu et al., 2021)	64.4	65.0
RaAS (Retriever)	56.1	57.5
RaAS (w/o Self-Training)	55.4	58.2
RaAS (Full Model)	66.1	65.3

Table 2: Overall results on GEOGRANNO and GEO(FunQL).

Baselines We compare our method with the following zero-shot/unsupervised baselines: 1) Cross-domain Zero Shot (Herzig and Berant, 2018), which is trained on other source domains and generalizes to target domains in OVERNIGHT and 2) GENOVERNIGHT (Wang et al., 2015), in which models are trained on synthesized ⟨CU, MR⟩ pairs; 3) SYNTH-SEQ2SEQ, in which the neural semantic parser is trained on the synthesized ⟨CU, MR⟩ pairs; 4) SSD (Wu et al., 2021), which use a paraphrase generation model to decode meaning representations. 5) AUTOQA (Xu et al., 2020), in which high-quality synthetic training data is generated by template-based data synthesizers and auto-paraphrasers.

Zero-shot Settings Any manual MR annotations are not required in our zero-shot settings. And, except for AutoQA, all of these zero-shot methods employ unannotated sentences as we do. We follow the hypothesis in GEOGRANNO: It is easy to access unlabeled utterances, which can typically be found in query logs, or generated by users experimenting with a prototype. Instead of unannotated sentences, AutoQA uses millions of generated sentences, which are not introduced in our method. AutoQA and our approach are two different strategies. The two methods are complementary, which means that our approach can be combined with AutoQA to eliminate the need for unannotated sentences.

3.1 Experimental Results

3.1.1 Overall Results

The overall results of different baselines and our method are shown in Table 1 and Table 2. We can see that:

1. **By exploiting the prior knowledge of PLMs and the sequences modeling ability of semantic parsers simultaneously, our RaAS framework**

		Bas.	Blo.	Cal.	Hou.	Pub.	Rec.	Res.	Soc.	Avg.
(1)	FULLMODEL	78.0	55.6	71.4	76.7	73.9	71.3	85.5	58.6	71.4
Inference										
(2)	(1) - Candidate Expansion	77.5	55.4	71.4	76.2	73.9	71.3	84.9	58.5	71.1
(3)	(1) - Retrieval Candidates	77.2	56.1	69.0	74.6	72.0	71.8	85.2	57.7	70.5
(4)	(3) - Reranking	75.7	56.6	65.5	73.0	70.1	72.7	85.2	57.6	69.6
(5)	(2) - Parser Scores	71.6	54.1	67.3	72.5	71.4	69.0	80.7	57.0	68.0
Prompt										
(6)	(1) - Prompt + Fine-Tuning	77.2	52.1	70.8	75.1	73.3	70.4	85.8	58.4	70.4
Self-Training										
(7)	(1) on Iter = 0	61.1	51.6	64.3	66.7	62.1	64.8	75.9	52.7	62.4
(8)	(1) on Iter = 1	75.4	54.1	70.8	75.1	72.0	70.8	85.5	59.0	70.3
(9)	(1) on Iter = 2	77.0	55.4	70.2	76.7	73.3	70.4	85.2	58.8	70.9
(10)	(1) on Iter = 4	77.5	55.6	70.8	76.7	73.9	71.3	85.2	58.3	71.2

Table 3: Ablation results of our model with different settings on OVERNIGHT.

achieves the best zero-shot semantic parsing performance. In all datasets, our method outperforms other baselines in the zero-shot settings, and further narrows the gap between zero-shot and supervised settings. These results demonstrate that zero-shot semantic parsers can be effectively constructed from the RaAS framework.

2. **The retrieval system can provide a good start without any annotated data.** Using pre-trained language models to retrieve meaning representations, the retrieval system can obtain an average accuracy rate close to 60% even without any supervision from manually labeled data. Considering the high recall rate of retrieval results, RaAS has the potential for later continuous improvement by ambiguous learning methods.

3. **Self-training can significantly improve the performances in all datasets.** In OVERNIGHT the average accuracy raises from 62.4% to 71.4%. As we mentioned before, the retrieval results have high recall rates but contain lots of noise. We think that the improvement of self-training comes mainly from candidate expansion and confidence re-estimation, which can establish global consistency gradually and reduce data noise iteratively.

3.1.2 Detailed Analysis

Self-training iterations In Table 3, Lines (7)-(10) show the accuracies on the test dataset as the number of iterations increases. We can see that: 1) The self-training protocol is effective. When we conduct more iterations, the performance gradually increases and stabilizes at a reasonable level – from 62.4% accuracy in Iter 0 to 71.4% in Iter 3 on OVERNIGHT. 2) The self-training process can reach its equilibrium within a few iterations, and the performance of RaAS can be stabilized around the third round.

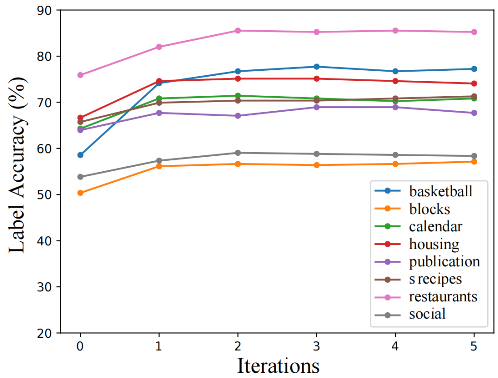

Figure 5: The accuracies on the validation set vary on the number of iterations in eight domains in OVERNIGHT.

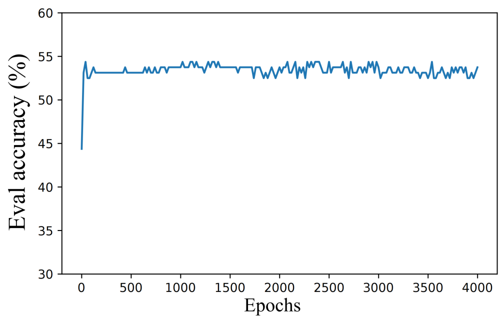

Figure 6: The accuracies on the validation set of Blocks domain in OVERNIGHT.

Composition of candidate set Line (2) in Table 3 shows the results of removing candidate expansion, where we only rerank retrieval candidates. Line (3) shows the results of removing retrieval candidates, where we only use beam search results of the current semantic parser.

1. **The effect of candidate expansion** If the candidate expansion is removed, the performances of RaAS decrease slightly. More importantly, during inferring, candidate expansion ensures the generation capability to produce various valid meaning representations, rather than only providing MRs in

the collected retrieval datastore.

2. **The effect of retrieval candidates** Without retrieval candidates, the performances drop slightly on average. We believe that this is because the beam search results are too similar, and the retrieval results can be a good supplement to them.

Reranking Line (4) in Table 3 shows the results of removing reranking, where we directly use beam search results of the semantic parser as output. The results of removing parser scores are shown in Line (5). We can see that without reranking, the average performance drops, but it still outperforms previous methods that exploit heavy data augmentations. However, without semantic parser scores, the performances will drop significantly.

The effect of prompt tuning Line (6) in Table 3 shows that, after changing the learning method to fine-tuning, the performances decrease slightly, which also proves the robustness and high generalization of prompt tuning.

The quality of confidence re-estimation In the Fig 5, we can see the accuracies on the validation set grow with the number of iterations. As the number of iterations increases, the performances gradually increase and stabilizes at a high level. This verifies that our self-training method can improve the quality of supervision signals iteratively by confidence re-estimation.

Few-shot settings The few-shot results are shown in Table 1. With the same few-shot settings as in previous studies, we employ T5-base to achieve comparable performances to T5-large and even T5-xl in previous work.

Training epochs Fig 6 shows the change of validation accuracies as the number of epochs increases. We can see that the performances of RaAS are stable, which verifies that our method is insensitive to the hyper-parameters of the number of training epochs in each iteration.

4 Related Work

Retrieval in Seq2Seq Tasks In semantic parsing, many previous studies (Su and Yan, 2017) have propose to employ paraphrase scores to retrieve or rerank MRs, which all follow the order of generating first and then scoring. Berant and Liang (2014) first generate a set of candidate MRs and choose the realization that best paraphrases the input. Yin and Neubig (2019) propose a set of reranking scorer for neural semantic parsers. Guo et al. (2019) combine a retrieval model and a meta-learner to employ the similar datapoints from the training data. Ren et al. (2020) construct parallel sentence pairs through retrieval, and conduct unsupervised machine translation models. Lu et al. (2021); Khandelwal et al. (2021); Parvez et al. (2021) enhance the representations of instances or the robustness of decoder by retrieval. Different from the common generate-then-score framework, the order of our RaAS framework is the reverse of them. We are the first to use retrieval results to obtain supervision for zero-shot semantic parsing.

Low Resource Semantic Parsing Many low resource semantic parsing methods have been proposed to reduce the demand for annotations(Artzi and Zettlemoyer, 2013; Sun et al., 2020; Sherborne and Lapata, 2022). Many weakly supervised learning are proposed (Berant et al., 2013; Reddy et al., 2014; Agrawal et al., 2019), such as denotation-based learning (Pasupat and Liang, 2016; Goldman et al., 2018), iterative searching (Dasigi et al., 2019). Semi-supervised semantic parsing is also proposed (Yin et al., 2018; Cao et al., 2019; Ye et al., 2019). One other strategy is to augment data. Wang et al. (2015) construct a semantic parsing dataset from grammar rules and crowd-sourcing paraphrase. Guo et al. (2018) produce pseudo-labeled data. Jia and Liang (2016) create new "recombinant" training examples with SCFG. Shin et al. (2021); Wu et al. (2021); Schucher et al. (2022) explore the training / decoding methods of PLMs for low-resource semantic parsing. Different from previous work, our framework focuses on obtaining and facilitating supervision signals rather than model design or data synthesization.

5 Conclusions

In this paper, we propose a novel method for zero-shot semantic parsing with a Retrieval as Ambiguous Supervision framework. We first retrieve the top-k similar meaning representations from the collected MR datastore. Then in self-training iterations, the candidates are employed to train parsers and refined by the candidate expansion and confidence re-estimation. We leverage the ambiguous supervision signal to train a prompt-based semantic parser and propose a confidence-driven self-training algorithm to refine the parser iteratively. The experiments show that the final semantic parser is greatly improved after iterative training.

Limitations

Firstly, due to the huge cost of large-scale PLMs, this paper only employs the T5-base as the backbone PLM in our experiments, therefore only limited analysis on the effect of model scale is presented. However, we believe a larger model will benefit our method by providing better language understanding and generation abilities.

Secondly, the synthesized canonical utterances need manually designed synchronous grammars, which are used to guide RaAS with knowledge about semantic representation language. Although most few-shot/zero-shot semantic parsing studies also rely on synchronous grammars, we leave how to model semantic representations without grammars as an open problem for future work.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. U1936207, 62122077 and 62106251. Furthermore, this research was supported by Meituan.

Ethics Consideration

This work presents RaAS, an effective framework for zero-shot semantic parsing. All of the involved datasets come from publicly available sources. The MRs and NLs are derived from several common public datasets (Kate et al., 2005; Wang et al., 2015; Herzig and Berant, 2019). The SCFGs are used for canonicalizing MRs, which are from OVERNIGHT and GEOGRANNO(Wang et al., 2015; Herzig and Berant, 2019). Pre-trained models and evaluation codes are all publicly accessible. The hyper-parameter settings are given in this paper. Our code and specification of dependencies will be released in the future.

References

Priyanka Agrawal, Ayushi Dalmia, Parag Jain, Abhishek Bansal, Ashish R. Mittal, and Karthik Sankaranarayanan. 2019. Unified semantic parsing with weak supervision. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4801–4810.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *TACL*, 1:49–62.

Anton Belyy, Chieh-yang Huang, Jacob Andreas, Emmanouil Antonios Platanios, Sam Thomson, Richard Shin, Subhro Roy, Aleksandr Nisnevich, Charles Chen, and Benjamin Van Durme. 2022. Guided k-best selection for semantic parsing annotation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 114–126. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1415–1425.

Ruisheng Cao, Su Zhu, Chen Liu, Jieyu Li, and Kai Yu. 2019. Semantic parsing with dual learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 51–64.

Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6806–6817. Association for Computational Linguistics.

Bo Chen, Le Sun, and Xianpei Han. 2018. Sequence-to-action: End-to-end semantic graph generation for semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 766–777.

Pradeep Dasigi, Matt Gardner, Shikhar Murty, Luke Zettlemoyer, and Eduard H. Hovy. 2019. Iterative search for weakly supervised semantic parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2669–2680.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Omer Goldman, Veronica Latcinnik, Ehud Nave, Amir Globerson, and Jonathan Berant. 2018. Weakly supervised semantic parsing with abstract examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1809–1819.

Daya Guo, Yibo Sun, Duyu Tang, Nan Duan, Jian Yin, Hong Chi, James Cao, Peng Chen, and Ming Zhou. 2018. Question generation from SQL queries improves neural semantic parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1597–1607.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2019. Coupling retrieval and meta-learning for context-dependent semantic parsing. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 855–866. Association for Computational Linguistics.

Jiaqi Guo, Qian Liu, Jian-Guang Lou, Zhenwen Li, Xueqing Liu, Tao Xie, and Ting Liu. 2020. Benchmarking meaning representations in neural semantic parsing. In *EMNLP*.

Jonathan Herzig and Jonathan Berant. 2018. Decoupling structure and lexicon for zero-shot semantic parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1619–1629.

Jonathan Herzig and Jonathan Berant. 2019. Don't paraphrase, detect! rapid and effective data collection for semantic parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3808–3818. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Zhanming Jie and Wei Lu. 2018. Dependency-based hybrid trees for semantic parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2431–2441.

Rohit J. Kate and Raymond J. Mooney. 2007. Learning language semantics from ambiguous supervision. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 895–900. AAAI Press.

Rohit J. Kate, Yuk Wah Wong, and Raymond J. Mooney. 2005. Learning to transform natural to formal languages. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 1062–1068.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Joohyun Kim and Raymond J. Mooney. 2010. Generative alignment and semantic parsing for learning from ambiguous supervision. In *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 543–551. Chinese Information Processing Society of China.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Wei Lu, Hwee Tou Ng, Wee Sun Lee, and Luke S. Zettlemoyer. 2008. A generative model for parsing natural language to meaning representations. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 783–792.

Xin Lu, Yijian Tian, Yanyan Zhao, and Bing Qin. 2021. Retrieve, discriminate and rewrite: A simple and effective framework for obtaining affective response in retrieval-based chatbots. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1956–1969. Association for Computational Linguistics.

Md. Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2719–2734. Association for Computational Linguistics.

Panupong Pasupat and Percy Liang. 2016. Inferring logical forms from denotations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Siva Reddy, Mirella Lapata, and Mark Steedman. 2014. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics*, 2:377–392.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2020. A retrieve-and-rewrite initialization method for unsupervised machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3498–3504. Association for Computational Linguistics.

Nathan Schucher, Siva Reddy, and Harm de Vries. 2022. The power of prompt tuning for low-resource semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 148–156. Association for Computational Linguistics.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4134–4153. Association for Computational Linguistics.

Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. pages 7699–7715.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yu Su and Xifeng Yan. 2017. Cross-domain semantic parsing via paraphrasing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1235–1246.

Yibo Sun, Duyu Tang, Nan Duan, Yeyun Gong, Xiaocheng Feng, Bing Qin, and Daxin Jiang. 2020. Neural semantic parsing in low-resource settings with back-translation and meta-learning. pages 8960–8967.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1332–1342.

Yuk Wah Wong and Raymond J. Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

Shan Wu, Bo Chen, Chunlei Xin, Xianpei Han, Le Sun, Weipeng Zhang, Jiansong Chen, Fan Yang, and Xunliang Cai. 2021. From paraphrasing to semantic parsing: Unsupervised semantic parsing via synchronous semantic decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5110–5121. Association for Computational Linguistics.

Silei Xu, Sina J. Semnani, Giovanni Campagna, and Monica S. Lam. 2020. Autoqa: From databases to QA semantic parsers with only synthetic training data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 422–434. Association for Computational Linguistics.

Hai Ye, Wenjie Li, and Lu Wang. 2019. Jointly learning semantic parser and natural language generator via dual information maximization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2090–2101.

Pengcheng Yin and Graham Neubig. 2019. Reranking for neural semantic parsing. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4553–4559.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. 2018. Structvae: Tree-structured latent variable models for semi-supervised semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 754–765.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI 96, IAAI 96, Portland, Oregon, USA, August 4-8, 1996, Volume 2.*, pages 1050–1055.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization.

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 11328–11339. PMLR.

Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5981–5990. IEEE.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

8

☑ B1. Did you cite the creators of artifacts you used?
8

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
7

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
7

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.