Semantic-aware Contrastive Learning for More Accurate Semantic Parsing

Shan Wu[1,2], **Chunlei Xin**[1,2], **Bo Chen**[1], **Xianpei Han**[1,3,*], **Le Sun**[1,3]

[1]Chinese Information Processing Laboratory [2]University of Chinese Academy of Sciences, Beijing, China [3]State Key Laboratory of Computer Science

Institute of Software, Chinese Academy of Sciences, Beijing, China

{wushan2018,chunlei2021,chenbo,xianpei,sunle}@iscas.ac.cn

Abstract

Since the meaning representations are detailed and accurate annotations which express fine-grained sequence-level semtantics, it is usually hard to train discriminative semantic parsers via Maximum Likelihood Estimation (MLE) in an autoregressive fashion. In this paper, we propose a semantic-aware contrastive learning algorithm, which can learn to distinguish fine-grained meaning representations and take the overall sequence-level semantic into consideration. Specifically, a multi-level online sampling algorithm is proposed to sample confusing and diverse instances. Three semantic-aware similarity functions are designed to accurately measure the distance between meaning representations as a whole. And a ranked contrastive loss is proposed to pull the representations of the semantic-identical instances together and push negative instances away. Experiments on two standard datasets show that our approach achieves significant improvements over MLE baselines and gets state-of-the-art performances by simply applying semantic-aware contrastive learning on a vanilla SEQ2SEQ model.

1 Introduction

Semantic parsing aims to translate natural language utterances into formal meaning representations(MRs), which has attracted much attention for many years (Wong and Mooney, 2007; Kate et al., 2005; Lu et al., 2008; Guo et al., 2019). Recent studies mostly treat semantic parsing as a neural sequence to sequence translation task via encoder-decoder frameworks (Dong and Lapata, 2016; Jia and Liang, 2016; Rabinovich et al., 2017; Chen et al., 2018; Zhao et al., 2020; Shao et al., 2020).

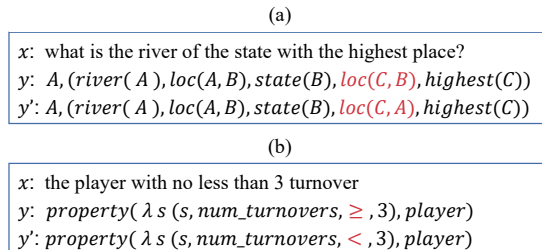To train neural semantic parsers, most studies employ Maximum Likelihood Estimation, which

(a)

x: what is the river of the state with the highest place?
y: $A, (river(A), loc(A, B), state(B), loc(C, B), highest(C))$
y': $A, (river(A), loc(A, B), state(B), loc(C, A), highest(C))$

(b)

x: the player with no less than 3 turnover
y: $property(\lambda s (s, num\_turnovers, \geq, 3), player)$
y': $property(\lambda s (s, num\_turnovers, <, 3), player)$

Figure 1: Meaning representations are detailed and accurate annotations which express fine-grained sequence-level semtantics. (a) the correct MR y and the predicted y' have similar token sequences but with very different semantics; (b) one token error (from $\geq$ to $<$) will reverse the semantics of MR.

optimizes the probabilities of the tokens in an autoregressive fashion. By decomposing sequence tasks into tokens, MLE training is good at n-gram based generation tasks such as machine translation and paraphrasing, but overlooks the semantics in a whole level.

Unfortunately, meaning representation is formal and detailed annotation, that is, it should be viewed as a sequence-level whole possessing fine-grained semantics. Such features make it hard to train accurate semantic parsers via MLE, which only computes loss token-by-token and is insensitive to small perturbations. For example, in Figure 1(b) even one token error (from $\geq$ to $<$) can reverse the semantics of a meaning representation: *Property (λ s (s num_turnovers $\geq$ 3), player)* to *Property (λ s (s num_turnovers $<$ 3), player)*. For a case of pp-attattchment problem, the MRs in Fig 1(a) have very similar token sequence but very different semantics. We analyze the error cases of a classical SEQ2SEQ parser in OVERNIGHT dataset[1], and found that the edit distances from 42.7% of the error parses to the correct MRs are only 1, and that of 74.2% of the error parses are $\leq$ 3. That is, most errors are due to the lack of ability to distinguish fine-grained semantics. There-

---

*Corresponding Author

[1]More statistical details can be found in Appendix A.1.

fore it is crucial to develop learning algorithms which can take the sequence semantics and the fine-granularity of meaning representations into consideration.

In this paper, we propose Semantic-aware Contrastive Learning (SemCL), which can learn semantic-aware, fine-grained meaning representations for accurate semantic parsing. To resolve the fine-granularity challenge, we sample negative instances in different divergence levels. And a multi-level online sampling algorithm is proposed to collect confusing and diverse instances. To resolve the sequence-level semantics challenge, we compare meaning representations as a whole, rather in token-by-token. Three semantic-aware similarity functions are designed to accurately measure the distance between utterances and meaning representations. Finally, we propose *ranked contrastive loss*, which is used to pull the representations of the semantic-identical instances together and push negative instances away (even if they look very similar to the positive ones). In this way, the semantic parsers can learn to distinguish fine-grained semantics and take the overall semantics into consideration.

In summary, the main contributions of this paper are:

- We propose a semantic-aware contrastive learning algorithm, which can effectively model the fine-grained and sequence-level semantics in semantic parsing. To our best knowledge, this is the first attempt to adopt contrastive learning for semantic parsing.

- We design an effective contrastive learning algorithm, which contains a multi-level online sampling algorithm, three semantic-aware similarity functions, and a ranked contrastive loss. This framework can also benefit other tasks which depend on the distinguishing ability of the fine-grained or whole-level semantics.

- Experiments on two standard datasets show that our approach achieves significant improvements over MLE baselines, and gets state-of-the-art performances.

2   Base SEQ2SEQ Parser

This paper uses the classical SEQ2SEQ semantic parser as our base model due to its simplicity and effectiveness (Dong and Lapata, 2016).

Encoder.   Given a sentence $\mathbf{x} = w_1, w_2, ..., w_n$, a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) or BERT (Devlin et al., 2019) can be used to map words into $\mathbf{h_x} = \mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n$.

Attention-based Decoder.   Given the sentence representation, the tokens of the logical forms are generated sequentially. Specifically, the decoder is first initialized with the hidden states of the encoder. Then at each step t, let $\phi(y_{t-1})$ be the vector of the previous predicted token, the current hidden state $\mathbf{s}_t$ is obtained from $\phi(y_{t-1})$ and $\mathbf{s}_{t-1}$. We calculate the attentioned source context representations for the current step t:

$$\alpha_t^i = \frac{\exp\left(\mathbf{s}_t \cdot \mathbf{h}_i\right)}{\sum_{i=1}^{n} \exp\left(\mathbf{s}_t \cdot \mathbf{h}_i\right)} \quad (1)$$

$$\mathbf{c}_t = \sum_{i=1}^{n} \alpha_t^i \mathbf{h}_i \quad (2)$$

and the next token is generated from the vocabulary distribution:

$$P(y_t|y_{<t}, \mathbf{x}) = \text{softmax}(\mathbf{W}_o[\mathbf{s}_t; \mathbf{c}_t] + \mathbf{b}_o) \quad (3)$$

MLE Learning.   The SEQ2SEQ model is trained by maximizing the likelihoods of the tokens in an autoregressive fashion:

$$\mathcal{L}_{seq2seq} = - \sum_{(\mathbf{x},\mathbf{y}) \in \mathbf{D}} \sum_{t=1}^{m} \log p(y_t|y_{<t}, \mathbf{x}) \quad (4)$$

where $\mathbf{D}$ is the corpus, $\mathbf{x}$ is the sentence, $\mathbf{y}$ is its logical form label.

However, such a token-by-token autoregressive training paradigm is insensitive to the overall semantics of the structured MR, making it hard to train effective and discriminative semantic parsers. We propose semantic-aware contrastive learning to help the semantic parsers to perceive the divergence of fine-grained semantics, which is overlooked in the existing autoregressive training approaches.

3   Semantic-aware Contrastive Learning for Semantic Parsing

In this section, we describe how to address the sequence-level semtantics and fine-granularity challenges via semantic-aware constrastive learning. The contrastive learning aims to disperse apart semantic-distinct instances and pull closer semantic-identical instances on vector representation space. Specifically, to learn to differentiate
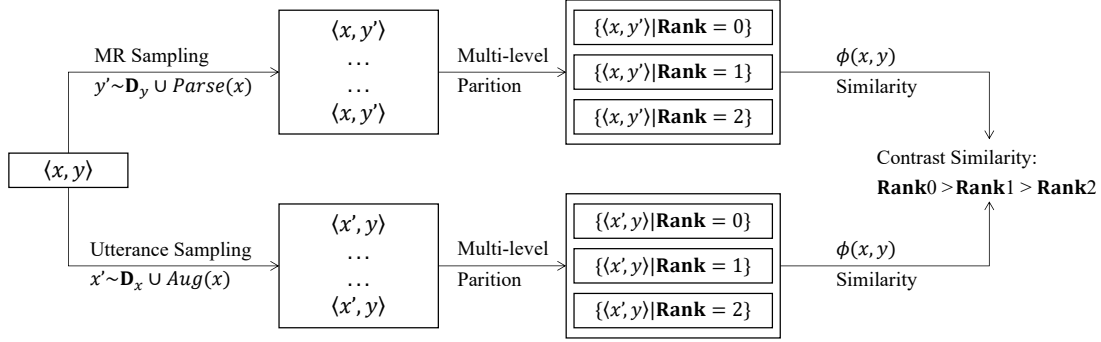
Figure 2: The architecture of our method, in which (1) we first collect samples via online sampling where $Parse(x)$ are the parses of the current model and $Aug(x)$ are the paraphrased augmentations; (2) then we divide samples into multi-levels, where $\mathbf{Rank} = 0$ indicates true positive, $\mathbf{Rank} = 2$ indicates true negative, and $\mathbf{Rank} = 1$ is vague instances containing paraphrases and potential aliases; (3) finally we use Ranked Contrastive Loss for optimization based on the similarities between the multi-level samples.

fine-grained meaning representations, we design a multi-level online sampling algorithm, which collects confusing negative samples and diverse positive samples in multi-level way. To comparing meaning representations as a whole at the sequence level, we design three semantics-aware compatibility functions. To learn accurate and discriminative semantic parsers, we propose *ranked contrastive loss* to support the multi-level samples. In following we describe them in detail.

3.1 Multi-level Online Instance Sampling

Contrastive learning algorithms learn good parameters by trying to pull positive instances closer and push negative instances away. Positive and negative instances play a fundamental role in constrastive learning (Karpukhin et al., 2020; Gao et al., 2021), and many studies focus on how to construct good positive and negative instances.

In semantic parsing, it is challenging to sample good contrastive instances. Firstly, because the meaning representation is formal and diverse, it is hard to tell the the changes on semantics after a small perturbation. Secondly, to distinguish the fine-grained semantic representations, contrastive learning needs accurate negative/positive samples. However, many instances are vague, which cannot be accurately categorized into positive-negative partitions. For example, paraphrasing, one common way to build positive instances, may changes the original fine-grained semantics. And two very different MRs may represent the same meaning, and cannot be treated as negative samples. To resolve the above challenges, we propose a multi-level partition algorithm to address vague instances, and sample instances via an online al-

gorithm.

3.1.1 Multi-Level Sample Partition

In contrastive learning, each instance is a pair of utterance and MR $\langle x, y \rangle$. To address the vagueness of instances, we divide samples into different levels according its confidence, and set each instance with a $\mathbf{Rank}$ value. Specifically, $\mathbf{Rank} = 0$ indicates true positive instances, $\mathbf{Rank} = 2$ indicates true negative instances, and $\mathbf{Rank} = 1$ indicates vague instances which may be correct. In following we describe how to divide instances into these levels and leave the sampling algorithm to next section.

$\mathbf{Rank} = 0$: This level contains true positive samples. We use the golden annotations in training corpus as positive instances. And two common types of aliases are also used as positive samples, which are show in Table 1. Given a MR, the utterance labeled as it and its aliases are used as positive samples.

$\mathbf{Rank} = 1$: This level contains vague samples which we cannot clearly identify whether it is into positive or negative. There are two types of vague instances. One is the utterance paraphrased ver-

Types	Examples
Expressions swapping in conjunction	$filter(type(meeting), attendee, =, concat(alice, bob)))$ $filter(type(meeting), attendee, =, concat(bob, alice)))$
Redundant type assurance	$filter(s, assists, <, ensureNumericEntity(3\ assist))$ $filter(s, assists, <, 3\ assist)$

Table 1: Two common types of MR aliases. As Guo et al. (2020) mentioned, because of the variety and composability of the aliases, there are a lot of diverse aliases of MR.

sion of instances, i.e., we paraphrase the annotated pair $\langle x, y \rangle$ and obtain $\langle x', y \rangle$. Because paraphrasing may change the original semantic, we set this instance as vague one. The other is the MR aliases version, i.e., for a positive instance $\langle x, y \rangle$ we view $\langle x, y'' \rangle$ and $\langle x'', y \rangle$ as vague instances if the annotated $\langle x'', y'' \rangle$ instance has the same execution result as y. Because the same execution result means they are potential aliases and may entail the same semantic, we use them as vague instances.

Because these samples are vague, directly adding them as positive will mislead the model, but ignoring them may reduce diversity of positive instances and thus affect the model generalization ability. Therefore we view them as vague instances.

Rank $= 2$: This level contains true negative instances. For utterance, negative MRs are the MRs with the wrong execution results. For MR, negative utterances are the ones labeled with the MRs producing wrong execution results .

3.1.2 Instance Sampling

There are two common sampling algorithms for contrastive learning:

1) Batch sampling: The positive and negative sample pairs are collected from the same batch. As shown in SimCLR(Chen et al., 2020), this algorithm is efficient and simple.

2) Online sampling: Given an annotated $\langle x, y \rangle$ pair, we sample its positive, vague, and negative instances during parsing. Given an input utterance, we use the top-K parses as candidates, and then devide them into **Rank** $0, 1, 2$ according to the methods described in above.

Because online sampling can collect hard negative samples(i.e., the top K ranked instances), this paper use it for better distinguishing confusing, fine-grained meaning representations.

3.2 Ranked Contrastive Loss

Traditional contrastive learning only considers positive and negative samples, and their instances are usually not fine-grained. In our semantic-aware contrastive learning, we need to deal with multi-level instances, and use special semantic-aware similarities.

To this end, we propose *ranked contrastive loss*, which aims to learn accurate and robust representations from the multi-level sample instances. Concretely, given sample instances in several lev-

els, ranked contrastive loss compare both utterances and meaning representations:

$$\mathcal{L}_{ctrs\_Utt} = \sum_{r=0,1} \sum_{\mathbf{Rank}_x = r} -log \frac{e^{\phi_\theta(x,y)/\tau}}{\sum_{\mathbf{Rank}_{x'} \geq r} e^{\phi_\theta(x',y)/\tau}}$$
(5)

$$\mathcal{L}_{ctrs\_MR} = \sum_{r=0,1} \sum_{\mathbf{Rank}_y = r} -log \frac{e^{\phi_\theta(x,y)/\tau}}{\sum_{\mathbf{Rank}_{y'} \geq r} e^{\phi_\theta(x,y')/\tau}}$$
(6)

, in which τ denotes a temperature parameter.

When there are only positive and negative samples as two ranks, the ranked contrastive loss can be gracefully degraded into the ordinary InfoNCE loss (van den Oord et al., 2018; Carse et al., 2021): With the minibatch $\mathcal{B}$ of size k, consisting of one positive pair (x, y) and $k-1$ negative pairs (x', y), the InfoNCE loss is defined as

$$\mathcal{L}_k = \mathbb{E}_{\mathcal{B}} \left[-\log \frac{e^{\phi_\theta(x,y)/\tau}}{e^{\phi_\theta(x,y)/\tau} + \sum_{i=1}^{k-1} e^{\phi_\theta(x_i',y)/\tau}} \right]$$
(7)

, which is also proved to be a lower bound on the mutual information of x and y.

The final training objective is to minimize the decoding loss and contrastive loss as follows:

$$\mathcal{L} = \mathcal{L}_{seq2seq} + \alpha \mathcal{L}_{ctrs\_Utt} + \beta \mathcal{L}_{ctrs\_MR} \quad (8)$$

where α and β are hyper-parameters that represent the weights of the contrastive learning. In this way, the model can reduce the influence of noise in sample instances and robustly improve the generalization by the augmentation of the instances.

3.3 Semantic-aware Compatibility Function

In contrastive learning, it is critical to measure the similarities between utterances and meaning representations, so that the positive $\langle x, y \rangle$ instances will have high similarity, and the negative instances will have low similarity. As described above, a semantic parsing system needs to measure the similarity by taking semantic representations as a whole. Specifically, we design three compatibility functions on sequence representations, attention-based representations and MR-conditioned representations.

Compatibility Function on Sequence Representations This similarity measure takes both utterance and meaning representations as two token se-

quences. We project the embedding representations of utterances and MRs onto the latent embedding space, and obtain the similarity between them:

$$\phi_{sr}(x,y) = mean(\mathbf{h_x})^T W_s \, mean(\mathbf{g_y}) \quad (9)$$

, in which $\mathbf{h_x}$ is the encoded contextual embedding of utterance $\mathbf{x}$ in SEQ2SEQ encoder, and $\mathbf{g_y} = \mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{|y|}$ is the encoded representations of y. An additional LSTM encoder is employed to represent MRs, which also shares the same token embeddings with the decoder.

Compatibility Function with Attention Because different tokens may have different importances, we extend the above mean pooling with attention mechanism as a soft selection to compute token-specific sentence representations.

$$a(\mathbf{h}_i, \mathbf{g}_t) = \frac{e^{\mathbf{h}_i^T W_a \mathbf{g}_t}}{\sum_{i'=1}^{|x|} e^{\mathbf{h}_{i'}^T W_a \mathbf{g}_t}} \quad (10)$$

$$\tilde{\mathbf{h}}_{y_t} = \sum_{i=1}^{|x|} a\left(\mathbf{h}_i, \mathbf{g}_t\right) \mathbf{h}_i \quad (11)$$

Then the compatibility function is:

$$\phi_{att}(x,y) = \sum_{t=1}^{|y|} \tilde{\mathbf{h}}_{y_t}^T W_{att} \mathbf{g}_t \quad (12)$$

MR-Conditioned Compatibility Function Semantic parsing is a SEQ2SEQ generation process, and the decoder decides which utterance tokens are used to decode a MR token y_t. Therefore we take these conditional association into consideration, and measure the similarity between x and y. In SEQ2SEQ decoding, c_t is the attentioned source context representation in the decoding step as in Equ. 2. Then the compatibility function is:

$$\phi_{cond}(x,y) = \sum_{t=1}^{|y|} c_t^T W_c \mathbf{g}_t \quad (13)$$

, where c_t captures the used parts of utterance representation in decoding.

4 Experiments

4.1 Experimental Settings

Datasets We conduct experiments on OVERNIGHT and GEOGRANNO, which involve various domains. Our implementations are public available[2].

OVERNIGHT This is a multi-domain dataset, which contains natural language queries paired with lambda DCS logical forms. The OVERNIGHT benchmark consists of eight semantic parsing datasets covering a range of semantic phenomena, which requires precise semantics learning ability to map natural language queries to the structured logical forms. We use the same train/test splits as Wang et al. (2015) to choose the best model during training.

GEOGRANNO This is an version of GEO (Zelle and Mooney, 1996; Herzig and Berant, 2019), which is labeled with lambda DCS logical forms. The dataset is constructed by paraphrases detecting. Crowd workers are employed to select the correct canonical utterance from candidate list. The generalization ability of models are requisite to handle 278 test queries from small numbers of train examples with only 487 instances. We follow the same splits in original paper (Herzig and Berant, 2019).

In all our experiments, the standard accuracy is used to evaluate systems. The accuracies on all datasets are calculated as the same as Jia and Liang (2016) and Herzig and Berant (2019).

Data Preprocessing Following Dong and Lapata (2016), we handle entities with Replacing mechanism, which replaces identified entities with their types and IDs. The entity mapping lexicons in Cao et al. (2019) are also used. The paraphrasing model is the trained paraphraser based on T5[3], and we paraphrase 20 different expressions for each utterances.

System Settings The bidirectional LSTM encoders are employed for utterances and MRs with 200 hidden units and 300-dimensional word vectors. We also use 200 hidden units and 300-dimensional word vectors for LSTM decoders. We initialize all parameters by uniformly sampling within the interval [-0.1, 0.1]. The batch size is set as 128. For each MR/utterance we collect 5 paraphrases and 100 random utterance/MR samples. In online sampling, after each training epoch, we collect the additional samples from the beam search results with the beam size as 20. We take $\alpha = 1$, $\beta = 1$, and $\tau = 0.3$. The first 5 epochs train the original model, and the following 25 epochs optimize the overall training loss. The beam size

	Bas.	Blo.	Cal.	Hou.	Pub.	Rec.	Res.	Soc.	Avg.
Previous									
SPO (Wang et al., 2015)	46.3	41.9	74.4	54.0	59.0	70.8	75.9	48.2	58.8
DSP-C (Xiao et al., 2016a)	80.5	55.6	75.0	61.9	75.8	-	80.1	80.0	72.7
RECOMB (Jia and Liang, 2016)	85.2	58.1	78.0	71.4	76.4	79.6	76.2	81.4	75.8
RECOMB (Jia and Liang, 2016)(+data)	87.5	60.2	81.0	72.5	78.3	81.0	79.5	79.6	77.5
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)	87.0	**66.2**	79.8	75.1	80.7	83.3	83.4	83.8	79.9
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
Our Methods									
SEQ2SEQ	86.4	61.4	73.8	68.3	76.4	77.8	78.3	82.8	75.7
SemCL (SR)	88.0	64.7	81.0	77.8	80.7	84.7	84.0	83.0	80.4
SemCL (Att)	87.7	65.2	81.5	76.2	80.7	82.4	83.7	83.8	80.2
SemCL (Cond)	**88.2**	65.7	**82.7**	**78.3**	**81.4**	83.8	**84.3**	**84.2**	**81.1**

Table 2: Overall results on OVERNIGHT.

of the decoder is set to 10. We use optimizer Adam(Kingma and Ba, 2015) with learning rate 0.001 for all experiments. In the main experiments and ablation experiments, our models use online sampling and multi-level partition by default.

4.2 Experimental Results

4.2.1 Overall Results

The overall results of baselines and different settings of our method are shown in Table 2 and Table 3. SR, Att, and Cond indicate the above three compatibility functions. We can see that:

1. **Semantic-aware contrastive learning is effective, which achieves state-of-the-art performance using a simple base model.** On OVERNIGHT and GEOGRANNO dataset, we both achieve state-of-the-art performance on average (81.1% and 73.4%). The results demonstrate the superiority of our contrastive learning algorithms.

2. **By taking the fine granularity and sequence-level semantics into consideration, the semantic-aware contrastive learning can significantly outperform MLE algorithm.** Compared with the MLE counterpart – SEQ2SEQ, the contrastive learning algorithm can lead to 5.4 and 1.8 accuracy improvements on OVERNIGHT and GEOGRANNO. This verifies that compared with MLE, our semantic-aware constrastive learning can learn more accurate semantic parsers.

3. **Semantic-aware similarity is critical for accurate semantic parsing.** We can see that, all compatibility functions show their advantages over MLE-baselines. And more accurate similarity measure can result better performance. Such as, MR-conditioned compatibility functions are

Previous Methods	
COPYNET (Herzig and Berant, 2019)	72.0
One-stage (Cao et al., 2020)	71.9
Two-stage (Cao et al., 2020)	71.6
SSD Word-Level (Wu et al., 2021)	72.9
SSD Grammar-Level (Wu et al., 2021)	72.0
Our Methods	
SEQ2SEQ	71.6
SemCL (SR)	73.0
SemCL (Att)	72.6
SemCL (Cond)	73.4

Table 3: Overall results on GEOGRANNO.

more stable in various domains and datasets. In general, the improvement of using semantic-aware contrastive learning is significant, regardless of which function in the three ones is used.

4.2.2 Detailed Analysis

Effect of multi-level sampling To analyze the effect of the multi-level samples partition, we conduct experiments with the positive-negative samples partition. The results are shown in Table 4. When there are only positive and negative samples, we try three ways to deal with the part of $\mathbf{Rank} = 1$: ignoring it or viewing it as positive or negative samples. We can see that treating it as negative is inadvisable, which brings significant performance degradation. We think this is because there are many positive examples in $\mathbf{Rank} = 1$ part, which should be gathered in representation space. And viewing them as negatives will mislead the model learning. We can see that treating it as positive also brings slight performance drop, which may be due to the noise samples. The results also show that our approach is better than neglecting it. We believe that ignoring them

	Bas.	Blo.	Cal.	Hou.	Pub.	Rec.	Res.	Soc.	Avg.
FULLMODEL	**88.2**	**65.7**	**82.7**	**78.3**	**81.4**	**83.8**	**84.3**	**84.2**	**81.1**
Multi-Level Samples									
$Rank = 1$ neglect	87.7	64.7	79.8	75.7	79.5	82.4	83.7	83.9	79.7
$Rank = 1$ as positive	**88.2**	63.7	81.0	75.1	78.3	81.0	80.4	82.9	78.8
$Rank = 1$ as negative	86.2	63.2	75.0	72.0	77.6	79.6	81.3	83.0	77.2
Precise Sampling									
Batch Sampling	87.5	63.7	80.4	75.1	78.3	82.9	82.5	82.6	79.1
Random Sampling	87.2	64.4	81.0	74.6	79.5	82.4	81.6	83.6	79.3
Contrastive Loss									
SEMCL ($\mathcal{L}_{ctrs\_MR}$ & SR)	87.5	62.2	81.0	71.4	77.0	82.4	80.4	82.8	78.1
SEMCL ($\mathcal{L}_{ctrs\_MR}$ & Att)	87.2	63.7	80.4	73.5	78.3	81.5	78.9	83.3	78.3
SEMCL ($\mathcal{L}_{ctrs\_MR}$ & Cond)	87.7	62.7	81.5	74.1	78.9	81.0	79.2	83.6	78.6
SEMCL ($\mathcal{L}_{ctrs\_Utt}$ & SR)	86.7	64.4	81.0	75.7	77.6	81.5	83.4	82.5	79.1
SEMCL ($\mathcal{L}_{ctrs\_Utt}$ & Att)	86.2	64.9	79.2	74.6	78.3	81.9	82.8	83.3	78.9
SEMCL ($\mathcal{L}_{ctrs\_Utt}$ & Cond)	87.0	65.2	81.0	76.2	78.9	81.5	83.1	83.4	79.5

Table 4: Ablation results of our model with different settings on OVERNIGHT.

(a) SEQ2SEQ (b) SemCL(without paraphrases) (c) SemCL(paraphrases as **Rank**=1)
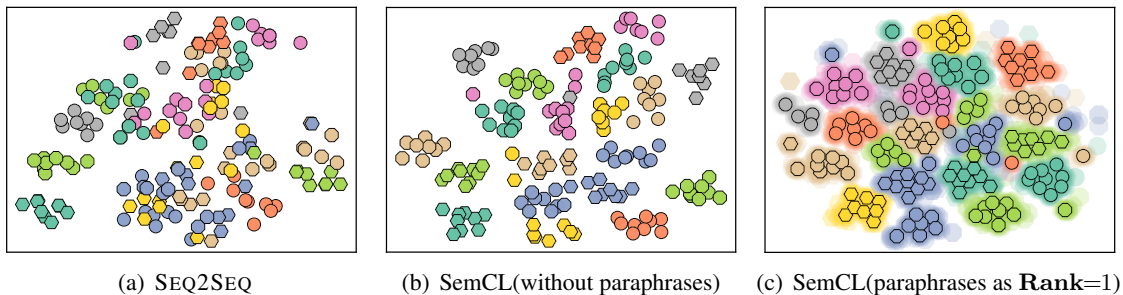
Figure 3: Comparison of t-SNE visualization for the learned utterance representations. The colors and markers indicate different MR labels. The transparent markers indicate the representations of the paraphrases.

will lead to insufficient generalization ability of the model. In general, it is problematic to employ vague instances in positive-negative partition, and our multi-level fashion can facilitate the learning of vague samples.

Effect of contrastive losses To investigate the effect of contrastive losses, we compare the settings with only contrastive learning on utterance side or MR side. The results in Table 4 show the performances of using contrastive losses on both sides are the best in all domains. We believe that by jointly optimizing the constrastive losses in both sides, the model can learn better sentence-side and semantic-side representations, which are beneficial to the model's awareness of fine-grained semantics.

Effect of instances sampling We conduct another experiment by changing sampling methods and the results are shown in Table 4: Batch Sampling denotes the sample instances are collected in the same batch; Random Sampling denotes for each utterance/MR we randomly select 100 MRs/utterances to form contrastive samples.

We can see the instance-level sampling is important. The Random Sampling on the instance-level is slightly better than Batch Sampling. We can see that the performances of Online Sampling (FULLMODEL) is significantly higher than other methods. This verifies that differentiating the hard negative samples, which are confusing on the current model, makes the model learn the fine-grained semantics more accurately.

Visualization of the representations space of utterances We use t-SNE (van der Maaten and Hinton, 2008) to visualize the utterance representations on a 2D map. The utterance representations are calculated by averaging the hidden states on words. The encoders of SEQ2SEQ models and our contrastive learning models(with sequence representations compatibility function) are used to obtain utterance representations. In Fig 3, we draw the most frequent 20 MRs and their utterances. Compared with SEQ2SEQ baseline, our two models learn smoother representation space, which explains why SemCL can yield better parsing performance. In Fig 3(c), we plot the 20

Input: articles published 2004 or more recent
Baselines: $listValue\,(filter\,(type\,(article), publication\_date,$ $=, cat\,(date\_2004, date\_2010)))$
SemCL: $listValue\,(filter\,(type\,(article), publication\_date,$ $\geq, date\_2004))$
(a) domain: PUBLICATION
Input: how many points did kobe bryant have when he had 3 block
Baselines: $listValue\,(property\,(filter\,(property\,(kobe\_bryant,$ $reverse(player)), num\_points, =, 3\,block), num\_points))$
SemCL: $listValue\,(property\,(filter\,(property\,(kobe\_bryant,$ $reverse(player)), num\_blocks, =, 3\,block), num\_points))$
(b) domain: BASKETBALL

Table 5: Cases on OVERNIGHT with MLE baseline and our SemCL. The MRs are simplified for readability.

paraphrases for each utterance with the transparent markers. Although there are some noise samples in the paraphrases, compared with no paraphrases, the generalization ability of the model is improved by our multi-level sampling mechanism.

Case study   In Table 5 we compare the parsed results of our model with that of the SEQ2SEQ baseline. In domain PUBLICATION, the utterance is mistaken by baseline for similar semantic "articles published in 2004 or 2010". Our model can well distinguish the fine-grained difference and generate the correct MR, even the expression about "recent" is rare in the training set. We think both generalization and discrimination of the models are improved. In domain BASKETBALL, the baseline over-translate the word "points". But SemCL can perceive the whole semantics and produce the right MR, which shows the effectiveness of treating both the utterance and the MR as a whole.

5   Related Work

Contrastive Learning.   Contrastive learning is a method of representation learning (Hadsell et al., 2006), which pulls the relevant embeddings together and pushes different ones apart to provide more effective representations (van den Oord et al., 2018; Chen et al., 2020). In NLP, contrastive learning is widely used in sentence representations learning (Qu et al., 2021; Gao et al., 2021; Kim et al., 2021; Giorgi et al., 2021). Contrastive learning also improves many natural language understanding tasks (Chen et al., 2021; Wang et al., 2021a,b; Qin et al., 2021). To the best of our knowledge, our work is the first attempt to adopt contrastive learning for semantic parsing.

Sequence-level Semantics.   In SEQ2SEQ tasks, many studies have been proposed to remedy the problem of MLE, such as, minimum risk training

(Shen et al., 2016), contrastive approaches (Lee et al., 2021) and reinforcement learning (He et al., 2016; Li et al., 2018). In semantic parsing, maximum marginal likelihood methodes are proposed to exploit consistent logical forms (Berant et al., 2013; Guu et al., 2017; Goldman et al., 2018). The structured learning methods are employed to maximize the margins or minimize the expected risks (Yih et al., 2015; Yin et al., 2018; Xiao et al., 2016b; Iyyer et al., 2017). Dual learning methods have been proposed, which also design and optimze rewards in sequence level (Cao et al., 2019, 2020). Globally normalized models are proposed to relieve the label bias problem in MLE (Huang et al., 2021). Different from previous works, our approach aims to acquire more discriminative sequence representations by contrastive learning.

Semantic Generalization.   Recently, generalization problem has become a research hotspot in semantic parsing. To generalize on various natural language expressions, semantic-invariance knowledge are introduced by paraphrasing (Berant and Liang, 2014; Wang et al., 2015; Dong et al., 2017; Herzig and Berant, 2019). There are also many studies for achieving generalization on meaning composition (Oren et al., 2020; Liu et al., 2020; Conklin et al., 2021; Bogin et al., 2021; Herzig et al., 2021). To generalize in low resources settings, data augmentation (Jia and Liang, 2016; Marzoev et al., 2020) and constrained decoding(Wu et al., 2021; Shin et al., 2021) methods are proposed. In this paper, the generalization of representations is improved by pulling the utterance and MRs with similar overall semantics closer in representation space.

6   Conclusions

This paper proposes Semantic-aware Contrastive Learning – an effective contrastive learning framework for semantic parsing, which takes the sequence-level semantics and the fine-granularity into consideration. Specifically, we propose a multi-level online sampling algorithm for accurately collecting confusing and diverse samples, and design three semantic-aware similarity functions to measure the similarity between utterances and MRs. We also propose Ranked Contrastive Loss to optimze the representations from the multi-level samples. Experimental results show that our approach can achieve state-of-the-art performance in several domains and datasets.

7 Limitations

There are two main limitations of this work. 1) Since additional negative instances are used for contrasting in our SemCL method, our method requires more training time than vanilla semantic parsing methods. 2) Our proposed method still relies on an amount of annotated data. Many contrastive learning methods have been proposed to resolve the low-resource tasks. We will exploit contrastive learning for low-resource semantic parsing in the future.

Acknowledgments

References

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. ACL.

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.

Ben Bogin, Sanjay Subramanian, Matt Gardner, and Jonathan Berant. 2021. Latent compositional representations improve systematic generalization in grounded question answering. *Trans. Assoc. Comput. Linguistics*, 9:195–210.

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.

Jacob Carse, Frank A. Carey, and Stephen J. McKenna. 2021. Unsupervised representation learning from pathology images with multi-directional contrastive predictive coding. In *18th IEEE International Symposium on Biomedical Imaging, ISBI 2021, Nice, France, April 13-16, 2021*, pages 1254–1258. IEEE.

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.

Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. 2021. CIL: contrastive instance learning framework for distantly supervised relation extraction. 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 6191–6200. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. 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 1597–1607. PMLR.

Henry Conklin, Bailin Wang, Kenny Smith, and Ivan Titov. 2021. Meta-learning to compositionally generalize. 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 3322–3335. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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*.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 875–886.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821.

John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. 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 879–895. Association for Computational Linguistics.

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. 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*.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. 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 4524–4535.

Kelvin Guu, Panupong Pasupat, Evan Zheran Liu, and Percy Liang. 2017. From language to programs: Bridging reinforcement learning and maximum marginal likelihood. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1051–1062. Association for Computational Linguistics.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828.

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.

Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. 2021. Unlocking compositional generalization in pretrained models using intermediate representations. *CoRR*, abs/2104.07478.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Chenyang Huang, Wei Yang, Yanshuai Cao, Osmar R. Zaïane, and Lili Mou. 2021. A globally normalized neural model for semantic parsing. *CoRR*, abs/2106.03376.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1821–1831. 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*.

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

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.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. 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 2528–2540. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3865–3878. Association for Computational Linguistics.

Qian Liu, Shengnan An, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Nanning Zheng, and Dongmei Zhang. 2020. Compositional generalization by learning analytical expressions. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. JMLR.

Alana Marzoev, Samuel Madden, M. Frans Kaashoek, Michael J. Cafarella, and Jacob Andreas. 2020. Unnatural language processing: Bridging the gap between synthetic and natural language data. *CoRR*, abs/2004.13645.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. 2020. Improving compositional generalization in semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2482–2495. Association for Computational Linguistics.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: improving entity and relation understanding for pre-trained language models via contrastive learning. 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 3350–3363. Association for Computational Linguistics.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2021. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Maxim Rabinovich, Mitchell Stern, and Dan Klein. 2017. Abstract syntax networks for code generation and semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1139–1149.

Bo Shao, Yeyun Gong, Weizhen Qi, Nan Duan, and Xiaola Lin. 2020. Multi-level alignment pretraining for multi-lingual semantic parsing. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3246–3256. International Committee on Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. 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*. The Association for Computer 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. *CoRR*, abs/2104.08768.

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.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021a. CLINE: contrastive learning with semantic negative examples for natural language understanding. 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 2332–2342. Association for Computational Linguistics.

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.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021b. CLEVE: contrastive pre-training for event extraction. 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 6283–6297. Association for Computational Linguistics.

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.

Chunyang Xiao, Marc Dymetman, and Claire Gardent. 2016a. Sequence-based structured prediction for 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*. The Association for Computer Linguistics.

Chunyang Xiao, Marc Dymetman, and Claire Gardent. 2016b. Sequence-based structured prediction for 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*. The Association for Computer Linguistics.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. 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 1321–1331. The Association for Computer Linguistics.

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.

Yuanyuan Zhao, Weiwei Sun, Junjie Cao, and Xiaojun Wan. 2020. Semantic parsing for english as a second language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6783–6794. Association for Computational Linguistics.

A Appendix

	Overall test cases	Error cases
Jaccard index	0.95	0.86
Token coverage	98.4%	90.1%
Edit distance	0.58	3.39

Table 6: Statistics of the distances between the predicted meaning representations and the correct ones. The error cases account for 13.6% of the overall test dataset.

A.1 Error Case Statistics

We use several indicators to measure the gap between the predicted meaning representation y' and the correct mentioned representation y. The basic SEQ2SEQ semantic parser is trained with MLE. The statistical results of BASKETBALL domain in OVERNIGHT are shown in Table 6. The three statistical indicators are used to represent the similarity between y' and y. The Jaccard index is calculated by:

$$J(y', y) = \frac{|y' \cap y|}{|y' \cup y|} \tag{14}$$

We can see that the model can easily produce outputs very similar to the correct results. Specifically, from 42.7% of the error parses to the correct MRs the edit distances are only 1, and from 74.2% of the error parses the edit distances of are ≤ 3. This reveals that the predicted wrong MRs of the parsers are very close to the correct answers, but it is still hard for the parsers to differentiate the fine-grained semantics precisely.