Unified Semantic Parsing with Weak Supervision

Priyanka Agrawal*, **Parag Jain, Ayushi Dalmia,**
Abhishek Bansal, Ashish Mittal, Karthik Sankaranarayanan
IBM Research AI
*pagrawal.ml@gmail.com
{pajain34, adalmi08, abbansal, arakeshk, kartsank}@in.ibm.com

Abstract

Semantic parsing over multiple knowledge bases enables a parser to exploit structural similarities of programs across the multiple domains. However, the fundamental challenge lies in obtaining high-quality annotations of (utterance, program) pairs across various domains needed for training such models. To overcome this, we propose a novel framework to build a unified multi-domain enabled semantic parser trained only with weak supervision (denotations). Weakly supervised training is particularly arduous as the program search space grows exponentially in a multi-domain setting. To solve this, we incorporate a multi-policy distillation mechanism in which we first train domain-specific semantic parsers (teachers) using weak supervision in the absence of the ground truth programs, followed by training a single unified parser (student) from the domain specific policies obtained from these teachers. The resultant semantic parser is not only compact but also generalizes better, and generates more accurate programs. It further does not require the user to provide a domain label while querying. On the standard OVERNIGHT dataset (containing multiple domains), we demonstrate that the proposed model improves performance by 20% in terms of denotation accuracy in comparison to baseline techniques.

1 Introduction

Semantic parsing is the task of converting natural language utterances into machine executable programs such as SQL, lambda logical form (Liang, 2013). This has been a classical area of research in natural language processing (NLP) with earlier works primarily utilizing rule based approaches (Woods, 1973) or grammar based approaches (Lafferty et al., 2001; Kwiatkowski et al., 2011; Zettlemoyer and Collins, 2005, 2007). Recently, there has been a surge in neural encoder-decoder techniques which are trained with input utterances and corresponding annotated output programs (Dong and Lapata, 2016; Jia and Liang, 2016). However, the performance of these strongly supervised methods is restricted by the size and the diversity of training data i.e. natural language utterances and their corresponding annotated logical forms. This has motivated the work on applying weak supervision based approaches (Clarke et al., 2010; Liang et al., 2017; Neelakantan et al., 2016; Chen et al., 2018), which use *denotations* i.e. the final answers obtained upon executing a program on the knowledge base and use REINFORCE (Williams, 1992; Norouzi et al., 2016), to guide the network to learn its semantic parsing policy (see Figure 3(a)). Another line of work (Goldman et al., 2018; Cheng and Lapata, 2018) is aimed towards improving the efficiency of weakly supervised parsers by applying a two-stage approach of first learning to generate program templates followed by exact program generation. It is important to note that this entire body of work on weakly supervised semantic parsing has

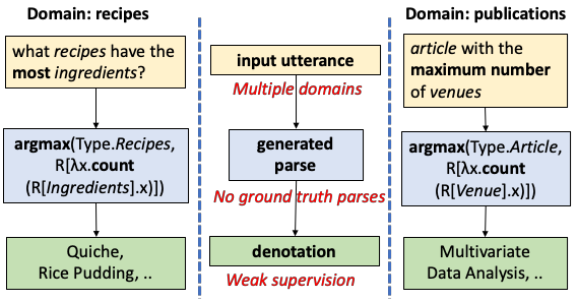


Figure 1: Examples for natural language utterances with linguistic variations in two different domains that share structural regularity (Source: OVERNIGHT dataset). Note that in this setup, we do not use ground truth parses for training the semantic parser.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4801–4810
Florence, Italy, July 28 - August 2, 2019.

been restricted to building a parser over a single domain only (i.e. single dataset).

Moving beyond single-domain to multiple domains, Herzig and Berant (2017) proposed semantic parsing networks trained by combining the datasets corresponding to multiple domains into a single pool. Consider the example in Figure 1 illustrating utterances from two domains, RECIPES and PUBLICATIONS, of the OVERNIGHT dataset. The utterances have linguistic variations *most* and *maximum number* corresponding to the shared program token *argmax*. This work shows that leveraging such structural similarities in language by combining these different domains leads to improved performance. However, as with many single-domain techniques, this work also requires strong supervision in the form of program annotations corresponding to the utterances. Obtaining such high quality annotations across multiple domains is challenging, thereby making it expensive to scale to newer domains.

To overcome these limitations, in this work, we focus on the problem of developing a semantic parser for multiple domains in the weak supervision setting using denotations. Note that, this combined multiple domain task clearly entails a large set of answers and complex search space in comparison to the individual domain tasks. Therefore, the existing multi-domain semantic parsing models (Herzig and Berant, 2017) fail when trained under weak supervision setting. See Section 6 for a detailed analysis.

To address this challenge, we propose a multi-policy distillation framework for multi-domain semantic parsing. This framework splits the training in the following two stages: 1) Learn domain experts (teacher) policy using weak supervision for each domain. This allows the individual models to focus on learning the semantic parsing policy for corresponding single domains; 2) Train a unified compressed semantic parser (student) using distillation from these expert policies. This enables the unified student to gain supervision from the above trained expert policies and thus, learn the shared semantic parsing policy for all the domains. This two-stage framework is inspired from policy distillation (Rusu et al., 2016) which transfers policy of a reinforcement learning (RL) agent to train a student network that is more compact and efficient. In our case, weakly supervised domain teachers serve as RL agents. For inference, only the compressed student model is used which takes as input the user utterance from any domain and outputs the corresponding parse program. It is important to note that, the domain identifier input is not required by our model. The generated program is then executed over the corresponding KB to retrieve denotations that are provided as responses to the user.

To the best of our knowledge, we are the first to propose a unified multiple-domain parsing framework which does not assume the availability of ground truth programs. Additionally, it allows inference to be multi-domain enabled and does not require user to provide domain identifiers corresponding to the input utterance. In summary, we make the following contributions:

- Build a unified neural framework to train a single semantic parser for multiple domains in the absence of ground truth parse programs. (Section 3)
- We show the effectiveness of multi-policy distillation in learning a semantic parser using independent weakly supervised experts for each domain. (Section 4)
- We perform an extensive experimental study in multiple domains to understand the efficacy of the proposed system against multiple baselines. We also study the effect of the availability of a small labeled corpus in the distillation setup. (Section 5)

2 Related Work

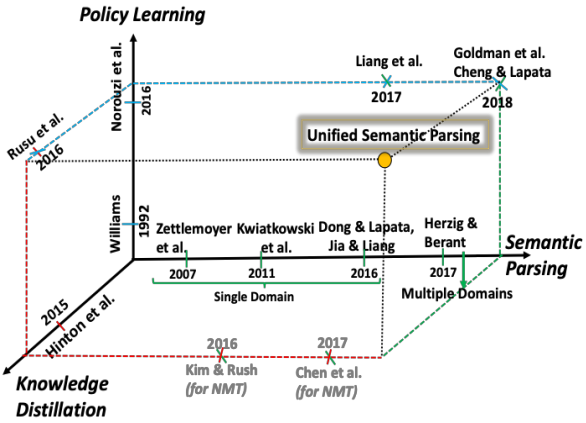


Figure 2: Illustration of the proposed work in the space of key related work in the area of semantic parsing, knowledge distillation and policy learning

This work is related to three different areas: semantic parsing, policy learning and knowledge

distillation. Figure 2 illustrates the placement of our proposed framework of unified semantic parsing in the space of the key related works done in each of these three areas. Semantic parsing has been an extensively studied problem, the first study dating back to Woods (1973). Much of the work has been towards exploiting annotated programs for natural language utterances to build single domain semantic parsers using various methods. Zettlemoyer and Collins (2007); Kwiatkowski et al. (2011) propose to learn the probabilistic categorical combination grammars, Kate et al. (2005) learn transformation from syntactic parse tree of natural language utterance to formal parse tree. Andreas et al. (2013) model the task of semantic parsing as machine translation. Recently, Dong and Lapata (2016) introduce the use of neural sequence-to-sequence models for the task of machine translation. Due to the cost of obtaining annotated programs, there has been an increasing interest in using weak supervision based methods (Clarke et al., 2010; Liang et al., 2017; Neelakantan et al., 2016; Chen et al., 2018; Goldman et al., 2018) which uses denotations, i.e. final answers obtained on executing a program on the knowledge base, for training.

The problem of semantic parsing has been primarily studied in a single domain setting employing supervised and weakly supervised techniques. However, the task of building a semantic parser in the multi-domain setting is relatively new. Herzig and Berant (2017) propose semantic parsing models using supervised learning in a multi-domain setup and is the closest to our work. However, none of the existing works inspect the problem of multi-domain semantic parsing in a weak supervision setting.

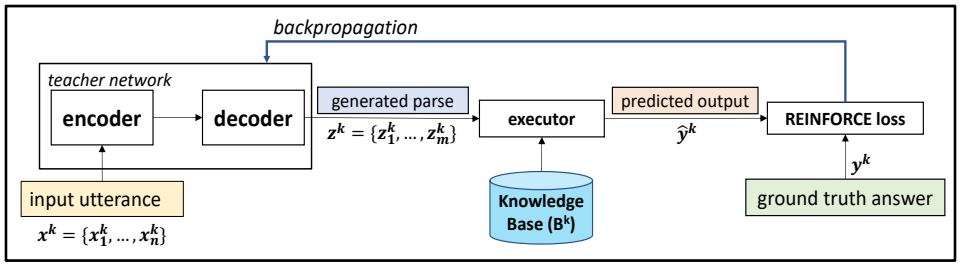
Knowledge distillation was first presented by Hinton et al. (2015) and has been popularly used for model compression of convolution neural networks in computer vision based tasks (Yu et al., 2017; Li et al., 2017). Kim and Rush (2016); Chen et al. (2017) applied knowledge distillation on recurrent neural networks for the task of machine translation and showed improved performance with a much compressed student network. Our proposed method of policy distillation was first introduced by Rusu et al. (2016) and is built on the principle of knowledge distillation and applied for reinforcement learning agents. Variants of the framework for policy distillations have also been proposed (Teh et al., 2017). To the best of our knowledge, our work is the first to apply policy distillation in a sequence-to-sequence learning task. We anticipate that the framework described in this paper can be applied to learn unified models for other tasks as well.

3 Proposed Framework

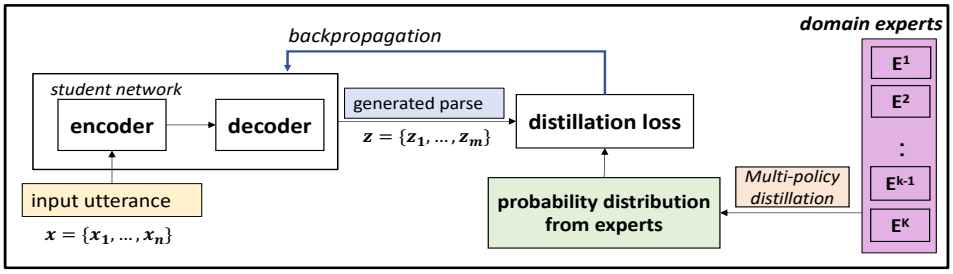
In this section, we first present a high level overview of the framework for the proposed unified semantic parsing using multi-policy distillation and then describe the models employed for each component of the framework.

We focus on the setting of 'K' domains each with an underlying knowledge-base $\mathbb{B}^1, \cdots, \mathbb{B}^K$. We have a training set of utterances X^k and the corresponding final denotations Y^k, for each domain $k \in 1, \cdots, K$. Unlike existing works (Herzig and Berant, 2017), we do not assume availability of ground truth programs corresponding to the utterances in the training data. Our goal is to learn a unified semantic parsing model which takes as input a user utterance $x_i^k = \{x_{i1}^k, \cdots, x_{in}^k\} \in X^k$ from any domain k and produces the corresponding program $z_i^k = \{z_{i1}^k, \cdots, z_{im}^k\}$ which when executed on the corresponding knowledge base $\mathbb{B}^k$ should return denotation $y_i^k \in Y^k$. In this setup, we only rely on the weak supervision from the final denotations Y^k for training this model. Moreover, the domain identifier k is not needed by this unified model.

We use multi-policy distillation framework for the task of learning a unified semantic parser. Figure 3 summarizes the proposed architecture. We first train parsing models (teachers) for each domain using weak supervision to learn domain-specific teacher policies. We use REINFORCE for training, similar to prior work on Neural Symbolic Machine (Liang et al., 2017) described briefly in Section 4.1. Next, we distill the learnt teacher policies to train a unified semantic parser enabled over multiple domain. (described in Section 4.2). Note that: (1) Our teachers are trained with weak supervision from denotations instead of actual parses and hence are weaker compared to completely supervised semantic parses. (2) State-of-the-art sequence distillation works (Kim and Rush, 2016; Chen et al., 2017) have focused on a single teacher-student setting.



(a) Domain specific expert policy E^k



(b) Learning a unified student S by distilling domain policies from experts $E^1, \cdots, E^K$

Figure 3: Proposed architecture diagram of unified semantic parsing framework. Figure 3(a) demonstrates the training of the experts E^k using weak supervision on the denotation corresponding to input utterance. Once we train all the domain experts $E^1, \cdots, E^K$ for the K domains, we use the probability distributions of the parse generated by these experts to train the student, thereby distilling the domain policies learnt by the teachers to the student as shown in Figure 3(b).

3.1 Model

In this section, we describe the architecture of semantic parsing model used for both teachers as well as the student networks. We use a standard sequence-to-sequence model (Sutskever et al., 2014) with attention similar to Dong and Lapata (2016) for this task. Each parsing model (the domain specific teachers $E^1, ..., E^K$ and the unified student S) is composed of an L-layer encoder LSTM (Hochreiter and Schmidhuber, 1997) for encoding the input utterances and an L-layer attention based decoder LSTM (Bahdanau et al., 2014) for producing the program sequences. Note that in this section, we omit the domain id superscript k.

Given a user utterance x, the aim of the semantic parsing model is to generate output program z which should ultimately result in the true denotations y. This user utterance $x = \{x_1, ..., x_n\}$ is input to the encoder which maps each word in the input sequence to the embedding $e = \{e_1, ..., e_n\}$ and uses this embedding to update its respective hidden states $h = \{h_1, ..., h_n\}$ using $h_t = \text{LSTM}(e_t, h_{t-1}; \theta_{enc})$, where θ_{enc} are the parameters of encoder LSTM. The last hidden state h_n is input to the decoder's first state. The decoder updates its hidden state s_t using $s_t = \text{LSTM}(c_{t-1}, s_{t-1}; \theta_{dec})$ where s_{t-1} is the embedding of output program token z_{t-1} at last step $t-1$ and θ_{dec} are the decoder LSTM parameters. The output program $\{z_1, ..., z_m\}$ is generated token-wise by applying softmax over the vocabulary weights derived by transforming the corresponding hidden state s.

Further, we employ beam search during decoding which generates a set of parses B for every utterance. At each decoding step t, a beam B_t containing partial parses of length t are maintained. The next step beam B_{t+1} are the $|B|$ highest scoring expansions of programs in the beam B_t.

4 Training

In this section we describe the training mechanism employed for the proposed multi-domain policy distillation framework for semantic parsing. The training process in our proposed framework has the following two components (Figure 3): (i) weakly supervised training for domain specific semantic parsing experts $E^1, ..., E^K$ and, (ii) distilling multiple domain policies to the unified student

S. We next describe each of these two components.

4.1 Domain-specific Semantic Parsing Policy

As described in the previous section, an individual domain specific semantic parsing model generates the program $z = \{z_1, ..., z_m\}$ which is executed on the knowledge base $\mathbb{B}$ to return the denotation $\hat{y}$. For brevity, we omit domain identifier k and instance id i in this section. In our setting, since labeled programs are not available for training, we use weak supervision from final denotations y similar to Liang et al. (2017) for each domain expert. As the execution of parse program is a non-differential operation on the KB, we use REINFORCE (Williams, 1992; Norouzi et al., 2016) for training which maximizes the expected reward. Reward $R(x, z)$ for prediction z on an input x is defined as the match score between the true denotations y for utterance x and the denotations obtained by executing the predicted program z. The overall objective to maximize the expected reward is as follows

$$\sum_x \mathbb{E}_{P_\theta(z|x)}[R(x,z)] = \sum_x \sum_z P_\theta(z|x)R(x,z) \approx \sum_x \sum_{z \in B} P_\theta(z|x)[R(x,z)]$$

where $\theta = (\theta_{enc}, \theta_{dec})$ are the policy parameters; B is the output beam containing top scoring programs (described in Section 3.1) and $P_\theta(z|x)$ is the likelihood of parse z

$$P_\theta(z|x) = \prod_t P_\theta(z_t|x, z_{1:t-1}) \quad (1)$$

To reduce the variance in gradient estimation we use baseline $b(x) = \frac{1}{|B|}\sum_{z \in B} R(x,z)$ i.e. the average reward for the beam corresponding to the input instance x. See Table 2 WEAKINDEP for the performance achieved for individual domains with this training objective.

Note that the primary challenge with this weakly supervised training is the sparsity in reward signal given the large search space leading to only a few predictions having a non-zero reward. This can be seen in the Table 2 WEAKCOMBINED when the entire set of domains is pooled into one, the numbers drop severely due to the exponential increase in the search space.

4.2 Unified Model for multiple domains

For the unified semantic parser, we use the same sequence-to-sequence model described in Section 3.1. The hyper-parameter settings vary from domain-specific models as detailed in Section 5.3. We use the multi-task policy distillation method of Rusu et al. (2016) to train this unified parser for multiple domains. The individual domain experts $E^1, ..., E^K$ are trained independently as described in Section 4.1. This distillation framework enables transfer of knowledge from experts $E^1, ..., E^K$ to a single student model S that operates as a multi-domain parser, even in the absence of any domain indicator with input utterance during the test phase. Each expert E^k provides a transformed training dataset to the student $D^k = \{(x_i^k, (\boldsymbol{p_\theta^k})_{\boldsymbol{i}})\}_{i=1}^{|X^k|}$, where $(\boldsymbol{p_\theta^k})_{\boldsymbol{i}}$ is the expert's probability distribution on the entire program space w.r.t input utterance x_i. Concretely, given m is the decoding sequence length and $\mathcal{V}$ is the vocabulary combined across domains, then $(\boldsymbol{p_\theta^k})_{\boldsymbol{i}} \in [0,1]^{m \times |\mathcal{V}|}$ denotes the expert E^k's respective probabilities that output token z_{ij} equals vocab token v, for all time steps $j \in \{1, \ldots, m\}$ and $\forall v \in \mathcal{V}$.

$$(\boldsymbol{p_\theta^k})_{\boldsymbol{i}} = \{\{p_\theta^k(z_{ij} = v; x_i^k, z_{i\{1:j-1\}})\}_{j=1}^{m}\}_{v=1}^{|\mathcal{V}|}$$

The student takes the probability outputs from the experts as the ground truth and is trained in a supervised manner to minimize the cross-entropy loss $\mathcal{L}$ w.r.t to teachers' probability distribution:

$$L(\theta^S; \theta^1, ..., \theta^K) = -\sum_{k=1}^{K}\sum_{i=1}^{|X^k|}\sum_{j=1}^{|m|}\sum_{v=1}^{|\mathcal{V}|} p_\theta^k(z_{ij} = v; x_i^k, z_{i\{1:j-1\}}) \log p_\theta^S(z_{ij} = v; x^k, z_{i\{1:j-1\}}) \quad (2)$$

where $\{\theta^k\}_{k=1}^K$ are the policy parameters of experts and θ^S are the student model parameters; similarly $p_\theta^S(z_{ij} = v; x^k, z_{i\{1:j-1\}})$ is the probability assigned to output token z_{ij} by student S. This training objective enables the unified parser to learn domain-specific parsing strategies from individual domains as well as leverage structural variations across domains. Therefore, the combined multi-domain policy S is refined and compressed during the distillation process thus rendering it to be more effective in parsing for each of the domains.

5 Experimental Setup

In this section, we provide details on the data and model used for the experimental analysis[1]. We further elaborate on the baselines used.

5.1 Data

We use the OVERNIGHT semantic parsing dataset (Wang et al., 2015) which contains multiple domains. Each domain has utterances (questions) and corresponding parses in $\lambda-$DCS form that are executable on domain specific knowledge base. Every domain is designed to focus on a specific linguistic phenomenon, for example, CALENDAR on temporal knowledge, BLOCKS on spatial queries. In this work, we use seven domains from the dataset as listed in Table 1.

We would like to highlight that we do not use the parses available in the dataset during the training of our unified semantic parser. Our weakly supervised setup uses denotations to navigate the program search space and learn the parsing policy. This search space is a function of decoder (program) length and vocabulary size. Originally, the parses have 45 tokens on an average with a combined vocabulary of 182 distinct tokens across the domains. To reduce the decoder search space, we normalize the data to have shortened parses with an average length of 11 tokens and 147 combined vocab size. We reduce the sequence length by using a set of template normalization functions and reduce the vocab size by masking named entities for each domain. An example of normalization function is the following: an entity utterance say of type *recipe* in the query is programmed by first creating a single valued list with the entity type i.e. `(en.recipe)` and then that property is extracted : `(call SW.getProperty ( call SW.singleton en.recipe ) ( string ! type ))` resulting in 14 tokens. We replace this complex phrasing by directly substituting the entity type under consideration i.e. `(en.recipe)` (1 token). Next, we show an example for a complete utterance: *what recipes posting date is at least the same as rice pudding*. Its original parse is:

```
(call SW.listValue (call SW.filter
(call SW.getProperty (call SW.singleton
 en.recipe) (string ! type)) (call
SW.ensureNumericProperty (string
posting_date)) (string >=)
(call SW.ensureNumericEntity (call
SW.getProperty en.recipe.rice_pudding
(string posting_date))))).
```

Our normalized query is *what recipes posting date is at least the same as e0*, where entity *rice pudding* is substituted by entity identifier *e0*. The normalized parse is as follows:

```
SW.filter en.recipe
 SW.ensureNumericProperty
    posting_date >=
        (SW.ensureNumericEntity
         SW.getProperty e0 posting_date)
```

It is important to note that this normalization function is reversible. During the test phase, we apply the reverse function to convert the normalized parses to original forms for computing the denotations. Table 1 shows the domain wise statistics of original and normalized data. It is important to note that this script is applicable for template reduction for any $\lambda-$DCS form.

We report hard denotation accuracy i.e. the proportion of questions for which the top prediction and ground truth programs yield the matching answer sets as the evaluation metric. For computing the rewards during training, we use soft denotation accuracy i.e. F1 score between predicted and ground truth answer sets.

Table 2 shows the accuracy with strongly supervised training (SUPERVISED). The average denotation accuracy (with beam width 1) of 70.6% which is comparable to state-of-the-art (Jia and Liang, 2016) denotation accuracy of 75.6% (with beam width 5). This additionally suggests that data normalization process does not alter the task complexity.

5.2 Baselines

In the absence of any work on multi-domain parser trained without ground truth programs, we compare the performance of the proposed unified framework against the following baselines:

1. **Independent Domain Experts** (WEAK-INDEPENDENT): These are the set of weakly supervised semantic parsers, trained independently for each domain using REINFORCE algorithm as described in Section 4.1. Note that these are the teachers in our multi-policy distillation framework.

2. **Combined Weakly Supervised Semantic Parser** (WEAK-COMBINED)): As per

[1]Code and data is available at https://github.com/pagrawal-ml/Unified-Semantic-Parsing

DOMAIN	ORIGINAL DATASET			NORMALIZED DATASET		
	UTTERANCE	PROGRAM		UTTERANCE	PROGRAM	
	Vocab	Vocab	Avg. Length	Vocab	Vocab	Avg. Length
BASKETBALL	340	65	48.3	332	58	20.5
BLOCKS	213	48	47.4	212	41	9.7
CALENDAR	206	54	43.7	191	46	8.8
HOUSING	302	58	42.7	293	48	8.5
PUBLICATIONS	190	44	46.2	187	38	8.5
RECIPES	247	49	42.6	241	40	7.8
RESTAURANTS	315	62	41.2	310	48	8.2
AVERAGE	259	54.3	44.6	252.3	45.6	10.3

Table 1: Training data statistics for original and normalized dataset. For each domain, we compare the #unique tokens (Vocab) in input utterances and corresponding programs; and average program length.

the recommendation in Herzig and Berant (2017), we pool all the domains datasets into one and train a single semantic parser with weak supervision.

3. **Independent Policy Distillation** (DISTILL-INDEPENDENT): We also experiment with independent policy distillation for each domain. The setup is similar to the one described in Section 4.2 used to learn K student parsing models, one for each individual domain. Each student model uses the respective expert model as the only teacher.

Following the above naming convention, we term our proposed framework as DISTILL-COMBINED. For the sake of completeness, we also compute the skyline SUPERVISED i.e. the sequence-to-sequence model described in Section 3.1 trained with ground truth parses.

5.3 Model Setting

We use the original train-test split provided in the dataset. We further split the training set of each domain into training (80%) and validation (20%) sets. We tune each hyperparameter by choosing the parameter from a range of values and choose the configuration with highest validation accuracy for each model. For each experiment we select from: beam width = {1, 5, 10, 20}, number of layers = {1,2,3,4}, rnn size for both encoder & decoder = {100, 200, 300}. For faster compute, we use the string match accuracy as the proxy to denotation reward. In our experiments, we found that combined model performs better with the number of layers set to 2 and RNN size set to 300 while individual models' accuracies did not increase with an increase in model capacity. This is intuitive as the combined model requires more capacity to learn multiple domains. Encoder and decoder maximum sequence lengths were set to 50 and 35 respectively. For all the models, RMSprop optimizer (Hinton et al.) was used with learning rate set to 0.001.

6 Results and Discussion

Table 2 summarizes our main experimental results. It shows that our proposed framework DISTILL-COMBINED clearly outperforms the three baselines WEAK-INDEPENDENT, WEAK-COMBINED, DISTILL-INDEPENDENT described in Section 5.2

Effect of Policy Distillation: DISTILL-INDEPENDENT are individual domain models trained through distillation of individual weakly supervised domain experts policies WEAK-INDEPENDENT. We observe that policy distillation of individual expert policies result in an average percentage increase of $\sim$ 10% in accuracy with a maximum of $\sim$ 33% increase in case of BLOCKS domains, which shows the effectiveness of the distillation method employed in our framework. Note that for CALENDAR domain, WEAK-INDEPENDENT is unable to learn the parsing policy probably due to the complexity of temporal utterances. Therefore, further distillation on the inaccurate policy leads to drop in performance. More systematic analysis on the failure cases is an interesting future direction.

Performance of Unified Semantic Parsing framework: The results show the proposed uni-

DOMAIN	WEAK-INDEPENDENT	WEAK-COMBINED	DISTILL-INDEPENDENT	DISTILL-COMBINED	SUPERVISED
BASKETBALL	33.8	0.5	33.8	**36.3**	81.0
BLOCKS	27.6	0.8	36.8	**37.1**	52.8
CALENDAR	**25.0**	0.6	12.5	17.3	72.0
HOUSING	33.3	2.1	42.3	**49.2**	66.1
PUBLICATIONS	42.2	6.2	45.9	**48.4**	68.3
RECIPES	45.8	2.3	61.5	**66.2**	80.5
RESTAURANTS	41.3	2.1	40.9	**45.2**	73.5
AVERAGE	35.5	2.1	39.1	**42.8**	70.6

Table 2: Test denotation accuracy for each domain comparing our proposed method DISTILLCOMBINED with the three baselines. We also report the skyline SUPERVISED.

fied semantic parser using multi-policy distillation (DISTILL-COMBINED) (as described in section 3) on an average has the highest performance in predicting programs under weak supervision setup. DISTILL-COMBINED approach leads to an increased performance by $\sim$ 20% on an average in comparison to individual domain specific teachers (WEAK-INDEPENDENT). We note maximum increase in the case of HOUSING domain with $\sim$ 47% increase in the denotation accuracy.

Effectiveness of Multi-Policy Distillation: Finally, we evaluate the effectiveness of the overall multi-policy distillation process in comparison to training a combined model with data merged from all the domains (WEAK-COMBINED) in the weak supervision setup. We observe that due to weak signal strength and enlarged search space from multiple domains, WEAK-COMBINED model performs poorly across domains. Thus, further reinforcing the need for the distillation process. As discussed earlier, the SUPERVISED model is trained using strong supervision from ground-truth parses and hence is not considered as a comparable baseline, rather a skyline, for our proposed model

6.1 Effect of Small Parallel Corpus

We show that our model can greatly benefit from the availability of a limited amount of parallel data where semantic parses are available. Figure 4 plots the performance of WEAK-INDEPENDENT and DISTILL-INDEPENDENT models for RECIPES domain when initialized with a pre-trained SUPERVISED model trained on 10% and 30% of parallel training data. As it can be seen, adding 10% parallel data brings an improvement of about 5 points, while increasing the parallel corpus size to only 30% we observe an improvement of about 11 points. The observed huge boost in performance is motivating given the availability of small amount of parallel corpus in most real world scenarios.

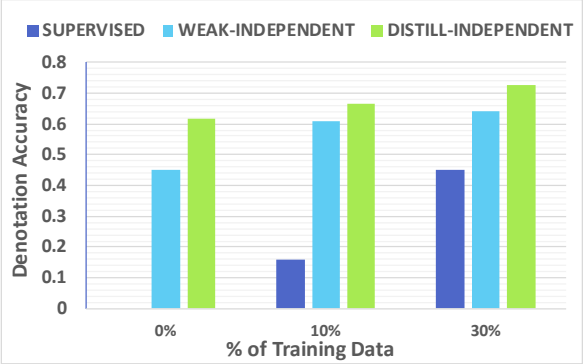


Figure 4: Effect of the fraction of training data on different models

7 Conclusions and Future Work

In this work, we addressed the challenge of training a semantic parser for multiple domains without strong supervision i.e. in the absence of ground truth programs corresponding to input utterances. We propose a novel unified neural framework using multi-policy distillation mechanism with two stages of training through weak supervision from denotations i.e. final answers corresponding to utterances. The resultant multi-domain semantic parser is compact and more precise as demonstrated on the OVERNIGHT dataset. We believe that this proposed framework has wide applicability to any sequence-to-sequence model.

We show that a small parallel corpus with annotated programs boosts the performance. We plan to explore if further fine-tuning using denotations

based training on the distilled model can lead to improvements in the unified parser. We also plan to investigate the possibility of augmenting the parallel corpus by bootstrapping from shared templates across domains. This would further make it feasible to perform transfer learning on a new domain. An interesting direction would be to enable domain experts to identify and actively request for program annotations given the knowledge shared by other domains. We would also like to explore if guiding the decoder through syntactical and domain-specific constraints helps in reducing the search space for the weakly supervised unified parser.

Acknowledgement

We thank Ghulam Ahmed Ansari and Miguel Ballesteros, our colleagues at IBM for discussions and suggestions which helped in shaping this paper.

References

Jacob Andreas, Andreas Vlachos, and Stephen Clark. 2013. Semantic parsing as machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 47–52.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473.

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 (Volume 1: Long Papers)*, pages 766–777. Association for Computational Linguistics.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935. Association for Computational Linguistics.

Jianpeng Cheng and Mirella Lapata. 2018. Weakly-supervised neural semantic parsing with a generative ranker. *CoRR*, abs/1808.07625.

James Clarke, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. 2010. Driving semantic parsing from the world's response. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 18–27. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 33–43. Association for Computational Linguistics.

O. Goldman, V. Latcinnik, U. Naveh, A. Globerson, and J. Berant. 2018. Weakly-supervised semantic parsing with abstract examples. In *Association for Computational Linguistics (ACL)*.

Jonathan Herzig and Jonathan Berant. 2017. Neural semantic parsing over multiple knowledge-bases. In *Association for Computational Linguistics*.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

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

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 (Volume 1: Long Papers)*, pages 12–22. Association for Computational Linguistics.

Rohit J. Kate, Yuk Wah Wong, and Raymond J. Mooney. 2005. Learning to transform natural to formal languages. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, AAAI'05, pages 1062–1068. AAAI Press.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327. Association for Computational Linguistics.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2011. Lexical generalization in ccg grammar induction for semantic parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1512–1523, Stroudsburg, PA, USA. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Jia Li. 2017. Learning from noisy labels with distillation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1928–1936.

Chen Liang, Jonathan Berant, Quoc V. Le, Ken Forbus, and Ni Lao. 2017. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23–33, Vancouver, Canada.

Percy Liang. 2013. Lambda dependency-based compositional semantics. *arXiv preprint arXiv:1309.4408*.

Arvind Neelakantan, Quoc V. Le, Martín Abadi, Andrew McCallum, and Dario Amodei. 2016. Learning a natural language interface with neural programmer. *CoRR*, abs/1611.08945.

Mohammad Norouzi, Samy Bengio, zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. Reward augmented maximum likelihood for neural structured prediction. In *Advances in Neural Information Processing Systems 29*, pages 1723–1731. Curran Associates, Inc.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. 2016. Policy distillation.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Yee Teh, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. 2017. Distral: Robust multitask reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4496–4506. Curran Associates, Inc.

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 (Volume 1: Long Papers)*, pages 1332–1342. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

W. A. Woods. 1973. Progress in natural language understanding: An application to lunar geology. In *Proceedings of the June 4-8, 1973, National Computer Conference and Exposition*, AFIPS '73, pages 441–450, New York, NY, USA. ACM.

Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. 2017. Visual relationship detection with internal and external linguistic knowledge distillation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1068–1076.

Luke Zettlemoyer and Michael Collins. 2007. Online learning of relaxed ccg grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, pages 658–666, Arlington, Virginia, United States. AUAI Press.