

Generating Synthetic Data for Task-Oriented Semantic Parsing with Hierarchical Representations

Ke Tran

Amazon Translate
Berlin, Germany
trnke@amazon.com

Ming Tan

Amazon Alexa
Cambridge, MA, USA
mingtan@amazon.com

Abstract

Modern conversational AI systems support natural language understanding for a wide variety of capabilities. While a majority of these tasks can be accomplished using a simple and flat representation of intents and slots, more sophisticated capabilities require complex hierarchical representations supported by semantic parsing. State-of-the-art semantic parsers are trained using supervised learning with data labeled according to a hierarchical schema which might be costly to obtain or not readily available for a new domain. In this work, we explore the possibility of generating synthetic data for neural semantic parsing using a pretrained denoising sequence-to-sequence model (*i.e.*, BART). Specifically, we first extract masked *templates* from the existing labeled utterances, and then fine-tune BART to generate synthetic utterances conditioning on the extracted templates. Finally, we use an auxiliary parser (AP) to filter the generated utterances. The AP guarantees the quality of the generated data. We show the potential of our approach when evaluating on the Facebook TOP dataset¹ for navigation domain.

1 Introduction

In this work, we investigate semantic parsing with hierarchical representations (Gupta et al., 2018) instead of the traditional logical forms (Zettlemoyer and Collins, 2005). Given an utterance x , our goal is to produce a tree-structured representation y of the utterance where additional information about intents and slots is introduced at the non-terminal nodes of the tree. We define a *template* z of a given annotation y as a result of replacing all terminal nodes by a generic [mask] node. Figure 1 shows an example of such an utterance x , its annotation y and the corresponding template z .

¹<http://fb.me/semanticparsingdialog>

```
x: What is the morning traffic hour
y: [IN:GET_INFO_TRAFFIC What is the
   [SL:DATE_TIME morning ] traffic hour ]
z: [IN:GET_INFO_TRAFFIC [mask]
   [SL:DATE_TIME [mask] ] [mask] ]
```

Figure 1: An example of an input utterance x , its desired output y , and the template z inferred from y . By definition, the template z above can be used to generate other utterances such as “how is the 5:00 traffic looking” or “Any construction on my morning route”.

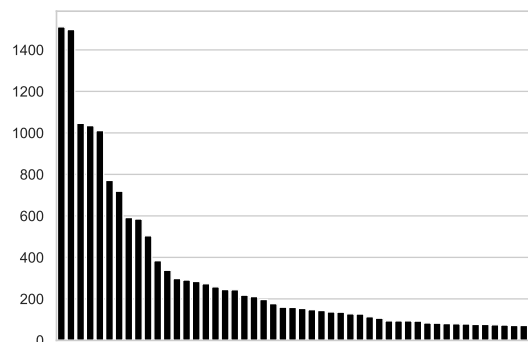


Figure 2: Frequency of most 50 common templates in Facebook TOP dataset. The frequency of z follows a power-law probability distribution.

The hierarchical representation for task-oriented parsing proposed in (Gupta et al., 2018) aims for ease of annotation and expressiveness. The dataset in their work, Facebook TOP, is the largest publicly available dataset in English for hierarchical semantic parsing. It has more than 44K annotated queries. We look at the distribution of the templates in Facebook TOP and found that the dataset is highly unbalanced (Figure 2). The 10 most frequent templates account for 30% of the training data and 14% of the data are singletons, which are utterances with only a single occurrence. This analysis suggests that it is beneficial to generate more synthetic data for templates with low frequencies. In the field of Natural Language Processing, us-

ing synthetic data via back-translation (Sennrich et al., 2016) has shown a great success for machine translation (Edunov et al., 2018). Unlike machine translation, generating synthetic data for hierarchical semantic parsing is less straightforward. Our work positions itself as one of the first to explore the possibility of generate text from graph (template) for semantic parsing.

In this paper, we propose a generic framework for augmenting a semantic parser with synthetic data. Our framework consists of two steps. First, we train a generator, followed by top- p sampling to generate diverse synthetic utterances conditioning on the above-mentioned templates. Generated utterances share similar hierarchical structures (*i.e.*, templates) with real training utterances while providing a wide spectrum of lexical variety. Second, we use an auxiliary parser for filtering on the generated candidates. The filtering step guarantees the quality of the synthetic data. Our generator is a sequence to sequence (seq2seq) model that is pre-trained on massive amount of monolingual data with text infilling objective (§2). We utilize BART (Lewis et al., 2020), a recently proposed denoising autoencoder, as our generator to avoid training it from scratch. The auxiliary parser can be arbitrary. We experiment with BART-based parser as well as state-of-the-art pointer network parser (s2s-pointer; Rongali et al., 2020).

The paper is structured as follows. We introduce our generative model for synthetic data in Section §2. Experimental results on Facebook TOP dataset and sub-sampled datasets to simulate low-resource scenario are presented in Section §3. Section §5 concludes the paper.

2 Denoising Sequence-to-Sequence as Generator

The generative story for generating synthetic data $\mathcal{Y}_{\text{syn}} = \{\tilde{y}_i\}_{i=1}^M$ is given by

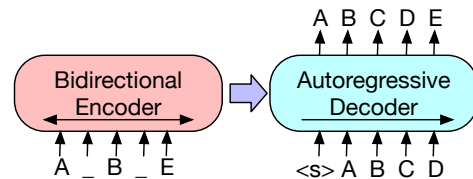
1. draw a template $z \sim p_\phi(z)$;²
2. draw an annotation $y \sim p_\theta(y|z)$ by filling each [mask] token in z by a word or sequence of words from vocabulary \mathcal{V} ;

Note that the transformation from annotation y to utterance x is deterministic by removing non-terminals from y . While $p_\phi(z)$ can be modeled by

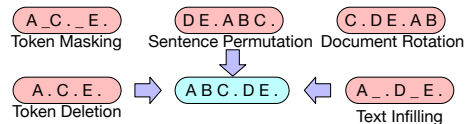
²During inference for generating synthetic data, we draw z uniformly in order to generate more annotations for templates in the long tail.

an autoregressive neural language model or a Probabilistic Context Free Grammar (Johnson, 1998), in this work we sample template z from seen templates in the data. We leave the possibility of generating new templates to future work.

We need a powerful conditional model $p_\theta(y|z)$ to generate annotation y . Thus, we choose BART, a pretrained denoising autoencoder for sequence-to-sequence, as our model. Figure 3a illustrates the idea behind BART. Given an input sequence (a stream of text), one of five types of noise (Figure 3b) is used to corrupt the input sequence. Then BART reconstructs the original sequence by maximizing the likelihood of the original sequence.



(a) BART is trained to reconstruct the corrupted input.



(b) Five different types of noise introduced in BART.

Figure 3: Overview of BART.

Since pretrained BART uses *text infilling* as noise to corrupt the input sequence, naturally we can use BART to infill the templates. Text infilling is the task where a number of spans in the original input sequence are replaced by a token [mask] and BART is trained to predict the replaced spans in the position of [mask] tokens. For our purpose of generating synthetic data, we fine-tune BART on an infilling dataset where the input is a template z with [mask] and the output is a linearized tree representation y where [mask] tokens are replaced by lexical words as shown in Figure 4.

BART source/target construction: We call out a few processing steps to construct this infilling dataset. First, non-terminal words are lowercased. We find this is necessary since the input will be tokenized by BART tokenizer and lowercasing non-terminal words prevents over-segmentation. Second, we make each of the closing brackets “]” in the original data explicit (*e.g.*, `in:get_distance`, `sl:destination`). This transformation provides the model explicit infor-

Source: [in:get_distance [mask] [sl:destination [in:get_location [sl:category_location [mask] sl:category_location] in:get_location] sl:destination] in:get_distance]
Target: [in:get_distance **How far is** [sl:destination [in:get_location [sl:category_location the coffee shop sl:category_location] in:get_location] sl:destination] in:get_distance]

Figure 4: Data for fine-tuning BART.

mation of the scope of the intents and slots.

Fine-tuning and generation: We fine-tune BART generator using the (*template, annotation*) pairs. After fine-tuning, we use the generator to generate full parse trees given templates. To increase the diversity of generated samples, we use top- p sampling (Holtzman et al., 2020) instead of beam search. The generator is trained to generate the tokenized labels together with the words. We remove generated annotations with invalid labels and convert the tokenized labels into the original tags in a post-processing step.

Auxiliary parser (AP) for filtering: In our preliminary experiments, we found that the generated samples are noisy. When we train our parser on the concatenation of both real and generated samples, the test accuracy degrades by 1.13% compared with a parser trained purely on real data. We therefore use an auxiliary parser (AP) to select robust samples. The filtering step is straightforward. First, we train an auxiliary semantic parser $f_{\theta}(\mathbf{x})$ on the original Facebook TOP dataset. We then use this trained AP to parse synthetic data $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)$ and keep those samples where the outputs of the parser $f_{\theta}(\tilde{\mathbf{x}}_i)$ match the synthetic labels $\tilde{\mathbf{y}}_i$ (i.e., $f_{\theta}(\tilde{\mathbf{x}}_i) = \tilde{\mathbf{y}}_i$). The AP for filtering can be different from the target parser we train for semantic parsing. Therefore, we propose three settings: (1) BART as AP and a sequence-to-sequence model with pointer networks (s2s-pointer; Rongali et al., 2020) as the target parser. (2) BART models for both AP and target parser. (3) s2s-pointer models for both AP and target parser. The comparisons and analysis are detailed in Section §3.

3 Experiments

We use Facebook TOP dataset in our experiments. Statistics of the dataset are shown in Table 1. While there are more than 31K annotated utterances in training data, the number of unique templates is about 6K. As we have shown in Section 1, the distribution of the templates is highly unbalanced. In training data, there are 1,511 annotations with the template [IN:UNSUPPORTED NAVIGATION [mask]] and 1,046 annotations with template

[IN:UNSUPPORTED [mask]]. In the case of no UNSUPPORTED setting (– UNSUPPORTED in Table 2), we exclude those annotations with UNSUPPORTED templates from train, valid, and test data. This results in 28,414 (template, annotation) pairs for training and 4,032 pairs for validation.

Condition	train	valid	test
+ UNSUPPORTED	31,279	4,462	9,042
– UNSUPPORTED	28,414	4,032	8,241

Table 1: Number of samples in Facebook TOP dataset with (+) and without (–) UNSUPPORTED utterances.

We fine-tune our BART generator using Adam optimizer (Kingma and Ba, 2015) with a linear warmup of 4,000 steps at the peak learning rate of $2e-5$. We pick the best model based on validation perplexity. After fine-tuning, we use the generator to sample 5 full parse trees per template.

The exact-match results for the three settings of using BART/s2s-pointer as auxiliary and target parser are given in Table 2. We first notice that the BART-based parser performs on-par with SOTA model based on pointer network and RoBERTa (Liu et al., 2019) feature extractor in the work of Rongali et al. (2020). This suggests that pretraining a general purpose seq2seq model is beneficial for downstream conditional generation task. We also see that using synthetic data brings additional 0.89% for BART-parser and 0.88% for s2s-pointer parser on the exact-match accuracy. The gain of using synthetic data is smaller when UNSUPPORTED utterances are present in training and testing data.

Table 3 shows the exact match accuracy of BART-based parser on testset with respect to template frequency f in training data. We see that synthetic data helps low-frequency templates ($f < 5$) the most (+1.36%). The gain of 0.67% for unseen templates ($f = 0$) suggests that there is a room for further improvement by generating new templates.

In order to support new domains (with new intents and slots) for the virtual assistants, we investigate the role of synthetic data when there is a little data available for the new domains. We sim-

Template:
[in:get_info_road_condition [mask] [sl:road_condition [mask]] [mask] [sl:path [mask]]]
Generate:
[in:get_info_road_condition is the road [sl:road_condition icy] on [sl:path I - 5]]
[in:get_info_road_condition Are the roads [sl:road_condition slick] on [sl:path I90]]
[in:get_info_road_condition Is there [sl:road_condition snow] on [sl:path the commute]]
[in:get_info_road_condition will the roads be [sl:road_condition slippery] on [sl:path my commute]]
[in:get_info_road_condition Are there any [sl:road_condition flooding] on [sl:path Route 66]]

Figure 5: Sample of five synthetic parse trees generated given a template. Colors indicate the corresponding generated spans per [mask] token. The data is reformatted for readability.

Data	AP filter	Target parser	– UNSUPPORTED		+ UNSUPPORTED	
			#Samples	Acc (%)	#Samples	Acc (%)
Real		BART	28,414	83.37	31,279	81.01
+ syn	BART	BART	53,679	84.26 (+0.89)	56,547	81.74 (+0.73)
Real		s2s-pointer	28,414	84.80	31,279	82.10
+syn	BART	s2s-pointer	53,679	85.31 (+0.51)	56,355	82.71 (+0.61)
+syn	s2s-pointer	s2s-pointer	89,629	85.68 (+0.88)	92,264	82.77 (+0.67)

Table 2: Exact-match results of our experiments. The AP filter can be a fine-tuned BART for parsing or a s2s-pointer model of Rongali et al. (2020)

Training data	$f \geq 5$	$f < 5$	$f = 0$
Real	89.46	74.70	61.90
+syn	90.30	76.06	62.57
Δ	0.84	1.36	0.67

Table 3: Exact-match accuracy on testset with respect to template frequency f in training data.

ulate this scenario by sub-sampling 6K utterances in the training data as follows: for each template in the training data, we randomly choose one utterance. We use this sub-sampled data for training our parser, generator, and AP. Table 4 shows the mean and variance of the accuracy on five random sub-sampled portions of the train data. We see that in this low resource setting, our approach boosts the accuracy by more than 2% absolute.

Training data	#Samples	Acc (%)
Real	6,000	72.24 \pm 0.05
+syn	30,000	74.31 \pm 0.05

Table 4: Average accuracy of five different runs for 6K training examples. The synthetic data is filtered by BART parser, which is trained on 6K samples.

4 Related Work

Using pretrained models to generate synthetic data has been studied recently (Amin-Nejad et al., 2020; Kumar et al., 2020). Their work however focuses on multi-class classification problems. Taking a step further, our work shows a viable path for structured output (*i.e.*, parse trees) problems.

5 Conclusions

We have proposed a novel approach for generating synthetic data for hierarchical semantic parsing. Our initial experiments show promising results of this approach and open up possibility for applying it to other problems with highly structured outputs in Natural Language Processing.

Acknowledgments

We thank reviewers for their constructive comments and suggestions. We also thank Raquel G. Alhama for proofreading this paper.

References

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. [Exploring transformer text generation for medical dataset augmentation](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. [Understanding back-translation at scale](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. [Semantic parsing for task oriented dialog using hierarchical representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *International Conference on Learning Representations*.
- Mark Johnson. 1998. Pcfg models of linguistic tree representations. *Comput. Linguist.*, 24(4):613–632.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.
- Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. [Data augmentation using pre-trained transformer models](#).
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. [Don’t parse, generate! a sequence to sequence architecture for task-oriented semantic parsing](#). In *Proceedings of The Web Conference 2020, WWW ’20*, page 2962–2968, New York, NY, USA. Association for Computing Machinery.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- 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*, page 658–666, Arlington, Virginia, USA. AUAI Press.