Unnatural Language Processing: Bridging the Gap Between Synthetic and Natural Language Data

Alana Marzoev
MIT CSAIL
marzoev@mit.edu

Jacob Andreas
MIT CSAIL
jda@mit.edu

Abstract

We study approaches for learning grounded interpretation of natural language utterances given only synthetically generated training examples. The effectiveness of synthetic-to-real transfer is investigated in both a standard sequence-to-sequence model and a new predictor defined by projection from the set of all utterances to those reachable by the synthetic generation procedure. These models are evaluated on a suite of semantic parsing tasks. On multiple tasks, a projection-based approach with no supervision from human annotators achieves accuracy within 5% of a state-of-the-art human-supervised approach.

1 Introduction

Data collection remains a major obstacle to the development of learned models for new language processing applications. Large text corpora are widely available for tasks like language modeling and machine translation [5, 4], but when building NLP systems that interact with the outside world in novel ways—whether through API calls (e.g. question answering) or physical actuators (e.g. robot instruction following)—it is rarely possible to use existing datasets. The element of interactivity precludes learning from text-only corpora, and effective models instead rely on training data from custom datasets constructed by skilled annotators. Acquiring these large, human-annotated corpora is an expensive and time-consuming undertaking. Indeed, human-centric data acquisition techniques are unscalable: sample efficiency poses a formidable challenge when learning to operate within interactive problem settings, since they are typically associated with combinatorial action spaces, and a large number of examples are typically required to learn even the simplest of tasks.

Existing solutions to this problem include replacing human users with learned models and allowing the input "language" to evolve freely during learning [10, 8]; or fine-tuning models initialized from language data on a downstream reward. In this paper, we instead propose to learn grounded meanings from *synthetic, naturalistic* language data in-domain, and use sentence representations learned from ungrounded data to generalize from synthetic utterances to real ones.

Our approach builds on a long but (within the machine learning community) largely disfavored body of work on hand-engineered grammars for language understanding [6]. The infrequent use of engineered grammars within larger learning systems is not without reason: while use of synthetic grammars drastically reduces dataset collection cost, such approaches are also brittle—the set of analyzable synthetic utterances is rarely a good approximation to the set of utterances produced by real language users. Therefore, models trained on "fake" data or constructed directly from hand-written grammars perform significantly worse when tested on real data. For example, in a dataset of instruction following tasks for drones, Blukis et al. report a test-time accuracy gap of 54% between models trained on synthetic data and those trained on real user utterances [2].

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

While transfer from synthetic problem domains to real ones has received significant attention in robotics and computer vision [12], the study of synthetic-to-real transfer has received comparatively little attention within NLP. One notable exception is semantic parsing, where prior work has attempted to reduce the mismatch between models trained on synthetic data and those trained on real data by making use of *supervised paraphrase data* (semantically equivalent restatements) to bootstrap semantic parsers in new application domains [1]. However, available paraphrase resources are limited and of varying quality, and the benefits of paraphrase-based semantic parsing over conventional approaches remains unclear.

In light of the recent success of self-supervised representation learning from text-only corpora, we believe it is time to revisit approaches to language learning in which synthetic data plays a central role. Instead of relying on paraphrases, we propose to use the *similarity of learned representations* as a bridge between synthetic data and real data, and present two approaches for learning grounded interpretations natural utterances using exclusively synthetic supervision.

In experiments on a set of standard question answering tasks [13], we show that a semantic parser can be successfully learned from only synthetic data and pretrained sentence representations. In three of eight problem domains studied, our approach achieves results comparable to a state-of-the-art semantic parser trained with human-generated utterances. These results have implications for both grounded language learning and emergent communication applications: while this work focuses on single-step prediction tasks, it is a first step towards training agents to coordinate via a scriptable, language-like protocol and then immediately generalize to interaction with real humans. More generally, the proposed approach provides a flexible strategy for automatically generating grounded datasets for new tasks.

2 Approach

Our goal is to learn a model $f : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is the space of of natural language inputs (e.g. questions or instructions) and $\mathcal{Y}$ is a space of outputs (e.g. semantic parses, action sequences, or dialogue responses). We assume access to an (arbitrarily large) dataset of synthetic training examples $\tilde{\mathcal{D}} = \{(\tilde{x}_i, \tilde{y}_i)\}$ for training f. Our goal is to minimize some expected loss $\mathbf{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x), y)]$ with respect to a "natural" data distribution $\mathcal{D}$.

We describe two models for unsupervised synthetic-to-real transfer, each of which can be augmented with pre-trained representations from an auxiliary task.

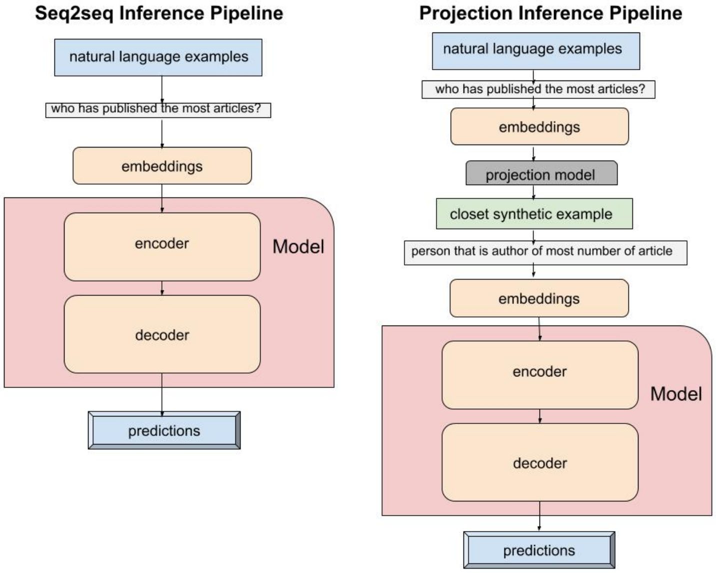
Seq2seq Inference Pipeline
natural language examples
who has published the most articles?
embeddings
encoder
decoder
Model
predictions
Projection Inference Pipeline
natural language examples
who has published the most articles?
embeddings
projection model
closet synthetic example
person that is author of most number of article
embeddings
encoder
decoder
Model
predictions

Figure 1: Sequence-to-sequence and projection model inference pipelines.

Sequence-to-sequence model The easiest approach is simply to train on the fake data and hope that the model will generalize to the real data as a result of overlap between $\tilde{\mathcal{D}}$ and $\mathcal{D}$. In this case we implement the model as

$$f(x) = \texttt{decode}(\texttt{encode}(\texttt{embed}(x))) \tag{1}$$

where embed produces an embedded representation of x (e.g. as a sequence of one-hot vectors), encode is a standard RNN encoder, decode is a task-appropriate decoder (e.g. an RNN for structured prediction tasks like semantic parsing, a classifier for sequential decisionmaking tasks like instruction following [3]) and f is optimized to minimize loss directly on $\tilde{\mathcal{D}}$.

Using pretrained representations: We expect the approach in Equation 1 to be an ineffective strategy for learning from scratch with synthetic data. However, incorporating a pretrained sentence representation model into embed and training as above corresponds to a standard fine-tuning procedure used by previous work for incorporating pretrained representations into a semantic parser. In this paper we incorporate pretrained representations by augmenting the one-hot word representations generated by embed with pretrained embeddings.

Projection model With or without pretraining, when evaluated on real data, f will be presented with inputs x from outside of the synthetic training distribution $\tilde{\mathcal{D}}$. In general, neural sequence prediction models offer no guarantee about how they will behave under shifts in the input distribution of this kind. The fact that the encoder model was trained with a great deal of data may not compensate for the fact that the decoder has only seen encodings of a small number of input sentences. How can we make use of robust, broad-coverage encoder outputs without exposing the decoder to out-of-distribution encodings?

Here we *project* from the natural data distribution $\mathcal{D}$ onto the set of synthetic examples $\tilde{\mathcal{D}}$ with respect to the distance function implied by the embedding model embed. That is, given a natural language input x, we first compute

$$\tilde{x}^* = \arg\min_{\tilde{x} \in \tilde{\mathcal{D}}} \ \delta(\texttt{embed}(x), \texttt{embed}(\tilde{x})) \tag{2}$$

under the similarity function δ, and finally predict

$$f(x) = \tilde{f}(\tilde{x}^*) \tag{3}$$

where $\tilde{f}$ is a model that can interpret synthetic data. This might a learned model (e.g. the sequence-to-sequence model above); under some circumstances, it may also be possible to compute $\tilde{x}$ directly by exploiting the structure of the synthetic generation procedure that produced $\tilde{\mathcal{D}}$ (e.g. when generating from a synchronous grammar).

Solving Equation 2 involves minimization over a combinatorial space of synthetic sentences. In this this paper, semantic domains are small enough that brute-force search is feasible, but future work might employ more sophisticated datastructures or learned search procedures to improve efficiency.

Using pretrained representations: note that in contrast to the sequence-to-sequence model, the projection step in Equation 2 does not involve a trained encoder or decoder model: any procedure for mapping strings to vectors will do. Any pretrained sentence representation can be incorporated directly into the representation produced by embed.

3 Experiments

As a first application of the proposed approach, we present experiments on a set of semantic parsing tasks. We report results on the Overnight datasets [13], each of which consists of a small set of compositional questions in a focused domain, designed to be representative of a rapid deployment scenario. The datasets consist of natural language utterances paired with logical forms. The logical representation of each domain is also equipped with a "canonical grammar" written by the authors of the datasets that can be used to jointly generate (string, logical form pairs), or to transduce from strings to logical forms and vice-versa. The canonical grammar is used to provide features for the original parser; here we use it both to generate the synthetic distribution $\tilde{\mathcal{D}}$ (consisting on the order of $100 - 1000$'s of examples) and to construct the synthetic predictor $\tilde{f}$.

	basketball	blocks	calendar	housing	publications	recipes	restaurants	social
Wang et al. [13]	.46	.42	.74	.54	.59	.71	.76	.48
seq-to-seq	.09	.17	.08	.11	.21	.26	.18	.08
seq-to-seq+BERT	.07	.21	.05	.13	.20	.16	.23	.14
projection	.37	.34	.30	.52	.43	.48	.52	.39
projection+BERT	.35	.44	.46	.50	.47	.47	.59	.46
% of supervised	.80	1.04	.62	.96	.80	.68	.78	.96

Table 1: Results for semantic parsing. The baseline approach of Wang et al. uses grounded, human-generated questions as supervision, while our seq-to-seq and projection models use only grounded synthetic data and pretrained word representations.

Model implementation details For the sequence-to-sequence model: the baseline model preprocesses sentences by downcasing and lemmatizing all words. `embed` represents each lemma with a learned representation. When incorporating pretrained representations, we use linear projections of contextual word embeddings trained using BERT [7]. In this case lemma representations are aligned to word piece boundaries [11] and concatenated to with the (un-preprocessed) BERT representation of the sentence. Projected embeddings are 256-dimensional, and both `encode` and `decode` are implemented as 1024-dimensional LSTMs [9].

For the projection model: sentences are lemmatized as in the sequence-to-sequence model; the vector sentence representation contains an indicator feature for each lemma. We again use BERT features from pretraining: here, we take the average contextual representation across all tokens in the sentence and concatenate this with the vector of lemma indicators. Cosine similarity is used as the distance function δ and $\tilde{f}$ is computed without training by inverting the synthetic grammar.

Results We report results on the test set for each of the Overnight datasets. As discussed, our approach makes use only of the canonical grammar in each domain and does not look at the training set. Predictions are evaluated on denotations rather than LF matches.

Results are presented in Table 1. The projection model is consistently better than the sequence-to-sequence model, and the addition of BERT features helps for most datasets. For three of the seven datasets, a projection-based model that makes no use of any natural training data is able to come within 5% of the accuracy of a state-of-the-art semantic parser trained on natural data.

3.1 Discussion

Our results on the semantic parsing benchmark demonstrate the promise of using projections as an approach to learning models that can better generalize from synthetic to natural data. Our method also has implications for best practices for training grounded models for new NLP applications—although pretraining cannot solve the problem of grounding natural language information into new domains and environments directly, early evidence suggests the possibility of a new paradigm for building language understanding systems. In this paradigm, key insights about the relationship between the world and language are explicitly encoded into the declarative synthetic data generation procedure rather than implicitly in the model's structure or through the use of a human-annotated dataset, and can take advantage of advances in machine learning and structured knowledge about human language.

In future work, we plan to evaluate our approach on sample-inefficient grounded reinforcement learning tasks, including instruction following with human-in-the-loop iterative corrections. We also hope to gain deeper practical insights into the extent of language coverage a synthetic grammar must provide to act as a starting point for synthetic to natural language transfer. Similarly, we are interested in better understanding how our approach performs on datasets in which the test sentences differ only from the training sentences by using words and phrases that are synonyms of things that appear in the synthetic dataset as opposed to when the test set contains known words reused in new configurations, and plan to investigate the types of mistakes that occur when new test examples involve phenomena entirely unrelated to any of the training data.

References

[1] Jonathan Berant and Percy Liang. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, 2014.

[2] Valts Blukis, Dipendra Misra, Ross A Knepper, and Yoav Artzi. Mapping navigation instructions to continuous control actions with position-visitation prediction. *arXiv preprint arXiv:1811.04179*, 2018.

[3] S.R.K. Branavan, Harr Chen, Luke S. Zettlemoyer, and Regina Barzilay. Reinforcement learning for mapping instructions to actions. In *ACL*, pages 82–90. Association for Computational Linguistics, 2009.

[4] Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F Zaidan. Findings of the 2011 workshop on statistical machine translation. In *WMT*. Association for Computational Linguistics, 2011.

[5] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

[6] Ann A Copestake and Dan Flickinger. An open source grammar development environment and broad-coverage english grammar using hpsg. In *LREC*, 2000.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] David Gaddy and Dan Klein. Pre-learning environment representations for data-efficient neural instruction following. In *ACL*, 2019.

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

[10] Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. Emergent translation in multi-agent communication. In *ICLR*, 2018.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[12] Eric Tzeng, Coline Devin, Judy Hoffman, Chelsea Finn, Xingchao Peng, Sergey Levine, Kate Saenko, and Trevor Darrell. Towards adapting deep visuomotor representations from simulated to real environments. In *WAFR*, 2016.

[13] Yushi Wang, Jonathan Berant, and Percy Liang. Building a semantic parser overnight. In *ACL*, 2015.