Semantic Parsing of Natural Language from Weakly Labeled Data

A Dissertation Submitted for the Degree of Doctor of Philosophy

Candidate: **LIU Qian**

Supervisor: **Professor ZHAO Qinping**
Associate Professor ZHOU Bin

School of Computer Science and Engineering

Beihang University, Beijing, China

Abstract

Semantic parsing over natural language is one of the most challenging directions in the field of natural language processing. It aims to build intelligent machines that can parse natural language questions from users into formal programs, and perform complex downstream tasks. On the one hand, semantic parsing is a major milestone on the road to understanding human language, and therefore has significant research value. On the other hand, semantic parsing allows users without any programming background to accomplish complex tasks with natural language, and thus has significant commercial value. However, program annotation in semantic parsing requires a lot of human and financial resources, which hinders the development of semantic parsing at scale, so it is necessary to carry out research on methods for semantic parsing over natural language under weak annotation. Weak annotation implies a weakening of both data quantity and data quality, which presents a great challenge for the research of semantic parsing. When the data quantity is relatively weak, the model faces the challenges of small program scale and small domain variety. When the data quality is relatively weak, the model faces the challenges of weal answer annotation and difficult conversation annotation construction.

Focusing on the fundamental topics of semantic parsing, we study and review the related works and research trends. And aiming at the challenging problems of small-scale program annotation, small-variety domain annotation, weak supervision of answer annotation and the difficulties of conversation annotation, we propose improving the program-oriented compositional generalization capabilities, improving the knowledge-base domain generalization capabilities, improving the performance of answer-driven semantic parsing model under weak supervision, and rapidly building conversational semantic parsing systems under semi-supervision. The main contributions of this dissertation are summarized as follows.

1. For the difficulty of small-scale program annotation, to enhance the compositional generalization ability of semantic parsing, we propose a memory-augmented neural model, which understand natural language sentences iteratively. The model consists of a Composer module for finding local spans of natural language, a Solver module for parsing natural language spans into program spans and a Memory module for storing variable values. These three modules cooperate to obtain the programs corresponding to natural language sentences. To address model optimization and alleviate the problem of reward sparsity, we propose to employ the hierarchical

reinforcement learning algorithm and the curriculum learning training strategy. Experiments on a typical compositional generalization benchmark show that our model effectively solves all the compositional generalization challenges proposed by previous work, with an accuracy of 100% on all test sets. The model is also the first neural approach that solves all previously proposed compositional generalization challenges.

2. For the difficulty of small-variety domain annotations, in order to enhance the generalization ability of semantic parsing models towards knowledge base domains, we propose a method which trains an entity linking model by erasing words in natural language sentences and a framework which combines entity linking with semantic parsing models. Without any extra labeling effort, our method can automatically generate pseudo-labels with the help of pre-training language models and semantic parsing datasets, and these pseudo-labels can be employed to learn an entity linking model. Regarding entity linking results as the prior, the decoder inside the semantic parser can generate programs more easily. Experimental results on four entity linking datasets show that the entity linking model using pseudo-labels as supervision outperforms baseline model by a large margin, achieving a performance improvement of 7.2%. Experimental results on two semantic parsing datasets for domain generalization show that our method can be flexibly applied to existing semantic parsers and significantly improve their domain generalization capabilities, achieving an absolute improvement of up to 9.8%.

3. For the difficulty of weak answer annotation, in order to enhance the model performance under weak supervision, we propose to first employ generative language models to generate answers for natural language questions, and then perform execution-guided pre-training. Different from the general pre-training methods which require crawling data from the Internet, our pre-training corpus can be synthesized automatically, thus allowing for both large scale and high quality. Experiments on three weakly supervised semantic parsing datasets show that the generative language model can effectively cope with the weakly supervised semantic parsing task when the task data is relatively large, achieving comparable performance to baselines. Additionally, the execution-guided pre-training method can boost the performance of the generative model significantly when the task data is relatively small, achieving an absolute improvement of up to 19.5%. Finally, our method achieves state-of-the-art results on all experimental datasets and can even be comparable with the baseline model under strong supervision.

4. For the difficulty of conversation annotations, in order to enhance the model performance under semi-supervision, we propose to decouple conversational semantic parsing into

two sub-tasks: conversation rewriting and single-round semantic parsing, which enables us to leverage existing single-round semantic parsing datasets. Furthermore, we collect the first conversation rewriting dataset for semantic parsing, FollowUp, which consists of 1, 000 conversations across 120 tables. To better accomplish the task of conversation rewriting, we propose a conversation rewriting method based on split-and-recombine, to better utilize the conversation flow by directly editing it. Experiments on the FollowUp dataset show that the conversation rewriting task is feasible for driving semi-supervised conversational semantic parsing, and the split-and-recombine approach can perform better than all baselines on improving semantic parsing. Finally, the conversational semantic parsing model trained by our method under semi-supervision can achieve 65% of the performance of full supervision.

Key words: Semantic Parsing, Weakly Supervised Semantic Parsing, Conversational Semantic Parsing, Compositional Generalization, Domain Generalization