Improving Tree-Structured Decoder Training for Code Generation via Mutual Learning

Binbin Xie\textsuperscript{1,2}, Jinsong Su\textsuperscript{1,2,*}, Yubin Ge\textsuperscript{3}, Xiang Li\textsuperscript{4}, Jianwei Cui\textsuperscript{4}, Junfeng Yao\textsuperscript{1} and Bin Wang\textsuperscript{4}

1. Xiamen University, Xiamen, China
2. Peng Cheng Laboratory, Shenzhen, China
3. University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
4. Xiaomi AI Lab, Beijing, China

\texttt{xdblb@stu.xmu.edu.cn, \{jssu, yao0010\}@xmu.edu.cn,}
\texttt{yubinge2@illinois.edu, \{lixiang21, cujianwei, wangbin11\}@xiaomi.com}

Abstract

Code generation aims to automatically generate a piece of code given an input natural language utterance. Currently, among dominant models, it is treated as a sequence-to-tree task, where a decoder outputs a sequence of actions corresponding to the pre-order traversal of an Abstract Syntax Tree. However, such a decoder only exploits the pre-order traversal based preceding actions, which are insufficient to ensure correct action predictions. In this paper, we first thoroughly analyze the context modeling difference between neural code generation models with different traversals based decodings (preorder traversal vs breadth-first traversal), and then propose to introduce a mutual learning framework to jointly train these models. Under this framework, we continuously enhance both two models via mutual distillation, which involves synchronous executions of two one-to-one knowledge transfers at each training step. More specifically, we alternately choose one model as the student and the other as its teacher, and require the student to fit the training data and the action prediction distributions of its teacher. By doing so, both models can fully absorb the knowledge from each other and thus could be improved simultaneously. Experimental results and in-depth analysis on several benchmark datasets demonstrate the effectiveness of our approach. We release our code at https://github.com/DeepLearnXMU/CGML.

Introduction

As an indispensable text generation task, code generation mainly focuses on automatically generating a snippet of code given a natural language (NL) utterance. Due to its great potential in facilitating software development and revolutionizing end-user programming, it has always been one of hot research topics in the communities of natural language processing and software engineering.

To achieve this goal, previous studies for code generation either require manually-designed grammar and lexicons (Zettlemoyer 2009; Zettlemoyer and Collins 2007) or exploit features for candidate logical forms ranking (Liang, Jordan, and Klein 2011). With the prosperity of deep learning in natural language processing, dominant models have now evolved into neural models, which possess the powerful capacity of feature learning and representation. Among them, the most prevalent one is the sequence-to-tree (Seq2Tree) models (Dong and Lapata 2016; Yin and Neubig 2017; Rabinovich, Stern, and Klein 2017; Dong and Lapata 2018; Shin et al. 2019; Sun et al. 2020), all of which are based on an encoder-decoder framework. Specifically, an encoder learns the word-level semantic representations of an input NL utterance, and then a tree-structured decoder outputs a sequence of actions, which corresponds to the pre-order traversal of an Abstract Syntax Tree (AST) and can be further converted into the code, 1) the structure of AST can be used to shrink the search space, ensuring the generation of well-formed code; and 2) the structural information of AST helps to model the information flow within the neural network, which naturally reflects the recursive structure of programming languages.

Despite the success of the above models, limited by the decoding manner based on the pre-order traversal, there still exists a serious defect in their decoders. Specifically, at each timestep, the current action can be only predicted from the vertical preceding actions. Since the dependencies of actions may come from other directions, i.e. the actions in horizontal direction, it may be insufficient to accurately predict the current action only with preceding actions in vertical direction. This happens especially when an important action locates in horizontal direction. Another possible case would be the strength of the dependencies on preceding actions turn to be weak with the increasing distances to those actions. Moreover, prediction errors in preceding actions sometimes even hurt the prediction of current action. Therefore, it is worth exploring the context in other direction to complement the current context of the pre-order traversal based decoder.

In this paper, we first explore the Seq2Tree model with breadth-first traversal based decoding, and then analyze the context modeling difference between different traversals based decodings (pre-order traversal vs. breath-first traversal). More importantly, we propose to introduce a mutual learning framework (Zhang et al. 2018b) to jointly train two Seq2Tree models with different traversals based de-
codings, both of which are expected to benefit each other. Our motivation stems from the fact that these two models are able to capture the context from different directions. That is, the conventional Seq2Tree model focuses on the pre-order traversal based preceding actions, while the other one mainly pays attention to the preceding actions based on breadth-first traversal. Hence, such difference brings the potential in enhancing both models.

Under the proposed framework, we continuously improve both models via mutual distillation, which involves synchronous executions of two one-to-one transfers at each training step. More specifically, we alternately choose one model as the student and the other as its teacher, and require the student to fit the training data and match action prediction distributions of its teacher. By doing so, both models can fully absorb the knowledge from each other and thus could be improved simultaneously.

The contributions of our work are summarized as follows:

- We explore the neural code generation model with breadth-first traversal and the mutual learning of models with different traversals based decodings can continuously enhance each other.
- We propose to introduce a mutual learning based model training framework for code generation, where Seq2Tree models with different traversals based decodings can continuously enhance each other.
- On several commonly-used datasets, we demonstrate the effectiveness and generality of our framework.

**Background**

In this work, we choose TRANX (Yin and Neubig 2018) as our basic model, which has been widely used due to its competitive performance (Yin and Neubig 2019; Shin et al. 2019; Xu et al. 2020). Note that both our explored decoder with breadth-first traversal and the mutual learning based model training framework are also applicable to other Seq2Tree models. In the following subsections, we first introduce how to convert NLs to code using TRANX, and then give a detailed description of the architecture of TRANX.

**Converting NLs to Code Using TRANX**

Typically, TRANX introduces ASTs as the intermediate meaning representations (MRs). Figure 1 gives an example of the NL utterance-to-code conversion for the Python code “if six.PY3: pass”. Concretely, such a conversion involves two stages: (1) the transition from the NL utterance into an AST. During this process, TRANX outputs a sequence of ASDL formalism based actions, corresponding to the AST in pre-order traversal; and (2) A user specified function AST-to-MR(*) is called to convert the generated AST into the code.

Specifically, at each timestep, one of three types of actions is evoked to expand the non-terminal node of the partial AST:

- **APPLYCONSTR[c]** actions apply a constructor c to the opening composite frontier field which has the same type as c, populating the opening node using the fields in c. If the frontier field has sequential cardinality, the action appends the constructor to the list of constructors held by the field.
- **RANX** takes an action \(a_{t-1}\) as its input. It is based on an attentional encoder-decoder framework, where a BiLSTM encoder is used to learn word-level semantic representations \(\{h_i\}_{i=1}^{|X|}\) of the input NL utterance \(X\), converted by an LSTM decoder into a sequence of tree-constructing actions.

Specifically, during the timestep \(t\) of decoding, its LSTM cell reads the embedding \(e(a_{t-1})\) of the previous action \(a_{t-1}\), the temporary hidden state \(s_{t-1}\), the vector \(p_t\) concatenated by the hidden state of the parent AST node and the type embedding of the current AST node, and the previous hidden state \(s_{t-1}\) to generate the hidden state vectors \(s_t\) and \(\tilde{s}_t\) as:

\[
s_t = f_{\text{LSTM}}(e(a_{t-1}); s_{t-1}; p_t, s_{t-1}),
\]

\[
\tilde{s}_t = \tanh(W_s [c_t; s_t]),
\]

where \(W_s\) is a parameter matrix and the attentional context vector \(c_t\) is produced from \(\{h_i\}\).

Finally, the probability of applying the action \(a_t\) is computed according to the type of the current node specified by the corresponding action of its parent node:

- **Primitive type.** In this case, TRANX takes an action \(a_t = \text{GENTOKEN}[v]\), where the token \(v\) can be either generated directly or copied from the input NL utterance as follows:

\[
p(a_t = \text{GENTOKEN}[v] | a_{<t}, x) = p(\text{gen}|a_{<t}, x) \cdot p_{\text{gen}}(v|a_{<t}, x) + (1 - p(\text{gen}|a_{<t}, x)) \cdot p_{\text{copy}}(v|x_{a_{<t}}, x).
\]

Here, we use three softmax functions based on \(s_t\) to the probability \(p(\text{gen}|a_{<t}, x)\) of choosing the generation operation, the probability \(p_{\text{gen}}(v|a_{<t}, x)\) of generating \(v\) and the probability \(p_{\text{copy}}(v = x_{a_{<t}}, x)\) of selecting to copy \(x_{a_{<t}}\), respectively.

- **Composite type.** On this condition, TRANX either applies an APPLYCONSTR action to produce the current node, or employs a REDUCE operation to terminate some node expansion from the AST parent node. Formally, the probability \(p(a_t = \text{APPLYCONSTR}\langle c \rangle | a_{<t}, x)\) of the action \(a_t\) is computed as \(\text{softmax}(e(a_t)^T W_a s_t)\), where \(W_a\) is the parameter matrix of the linear mapping.

**TRANX**

TRANX is based on an attentional encoder-decoder framework, where a BiLSTM encoder is used to learn word-level semantic representations \(\{h_i\}_{i=1}^{|X|}\) of the input NL utterance \(X\), converted by an LSTM decoder into a sequence of tree-constructing actions.
Training Objective. Given a training corpus $D = \{ (x, a) \}$, the training objective $J_{\text{MLE}}(D; \theta)$ is defined as

$$J_{\text{MLE}}(D; \theta) = \sum_{(x,a) \in D} J_{\text{MLE}}(x, a; \theta),$$

(4)

$$J_{\text{MLE}}(x, a; \theta) = -\frac{1}{T} \sum_{t=1}^{T} \log p(a_t|a_{<t}, x; \theta).$$

(5)

Our Approach

In this section, we first briefly describe two Seq2Tree models with different traversals (pre-order traversal and breadth-first traversal) based decodings. Then, we propose to introduce mutual learning (Zhang et al. 2018b) based model training framework into code generation, where these two models can be enhanced simultaneously.

Seq2Tree Models

TRANX. We omit the description of TRANX, which has been provided in Background section.}

TRANX’. As a variant of TRANX, TRANX’ provides a new perspective of code generation, where the AST is generated via breadth-first traversal and thus the AST context in horizontal direction can be leveraged for action predictions.

To train TRANX’, we traverse the AST of each training example in breadth-first order to generate a reference action sequence $a'$. As shown in the middle of Figure 1, for the code “if six.PY3: pass”, we traverse its AST in the breadth-first order and then reorganize the corresponding actions into an action sequence: $a' = a'_1, a'_2, \ldots, a'_8$. Note that although reference action sequences of TRANX and TRANX’ are different, they are from the same AST.

Likewise, we then parameterize the prediction probability using a LSTM network: $p(a'|x) = \prod_{t'=1}^{T} p(a'_{t'}|a'_{<t'}, x)$. The main equations of this LSTM are almost the same as those of TRANX, the only difference is that the definitions of preceding actions are different. Very importantly, $e(a_{t-1})$ of Equation 1 is replaced with $e(a'_{t-1})$.

Apparently, during different traversals based decodings, TRANX and TRANX’ focus on the AST-based context from different directions. For example, for the action prediction
Mutual Learning Based Model Training

As illustrated in Figure 2, we then propose to introduce a mutual learning framework to jointly train \textsc{TransX} and \textsc{TransX}'. Need to add that our framework is also suitable for joint training of more than two models.

In order to facilitate the understanding of our framework, we also depict its training procedure in Algorithm 1. Using this framework, we continuously enhance these two models via mutual distillation, where two one-to-one knowledge transfers in reverse directions are synchronously executed at each training step (Lines 6-15). During the one-to-one transfer procedure, we expect each model to fully learn knowledge from not only its own training data but also the other.

To this end, we require each considered model to fit the training data and match the action prediction distributions of the other model. In addition to the conventional MLE training data and match the action prediction distributions of the other one. Suppose \( \theta \) and \( \theta' \) to be the parameter sets of \textsc{TransX} and \textsc{TransX}', we define the following objective function to update their parameters:

\[
J(D, \theta) = \sum_{(x,a) \in D} \{ J_{\text{MLE}}(x, a; \theta) + \}
\]

\[
\frac{\lambda}{T} \sum_{n \in a} \text{KL}(p(a^t_n | a_{t < t(n)} \backslash n, x; \theta') \| p(a_{t(n)} | a_{t < t(n)} \backslash n, x; \theta)) + \}
\]

\[
J(D, \theta') = \sum_{(x,a) \in D} \{ J_{\text{MLE}}(x, a'; \theta') + \}
\]

\[
\frac{\lambda}{T} \sum_{n \in a} \text{KL}(p(a^t_n | a_{t < t(n)} \backslash n, x; \theta') \| p(a_{t(n)} | a_{t < t(n)} \backslash n, x; \theta')) \}
\]

where \( t(n) \) and \( t'(n) \) denote the timestamps of the AST node \( n \) during the pre-order and breadth-first traversals of the AST, respectively. \text{KL}(-||-) is the KL divergence, and \( \lambda \) is the coefficient used to control the impacts of different losses. Note that both our KL terms act on the AST node-level outputs of different models, that is, \( p(a^t_n | a_{t < t(n)} \backslash n, x; \theta') \) and \( p(a_{t(n)} | a_{t < t(n)} \backslash n, x; \theta) \), where the AST node \( n \) may correspond to different timestamps of our two models. Back to Figure 1, the node \( n_6 \) corresponds to the 6-th and 3-rd timesteps of pre-order and breadth-first traversals, respectively.

We repeat the above-mentioned knowledge transfer process, until both models converge.

**Algorithm 1** The training procedure of our framework.

**Input:** Training set \( D \), validation set \( D_v \).

1: repeat
2: \textbf{repeat}
3: \quad Load a batch size of instances \( B \in D \)
4: \quad \( J(B, \theta) = 0 \)
5: \quad \( J(B, \theta') = 0 \)
6: \quad \textbf{for} each instance \( (x,a) \in B \) \textbf{do}
7: \quad \textbf{end for}
8: \quad Calculate the loss \( J(x,a; \theta) \) of \textsc{TransX}
9: \quad Calculate the loss \( J(x,a'; \theta') \) of \textsc{TransX}'
10: \quad \( J(B, \theta') := J(x,a; \theta') \)
11: \quad \( J(B, \theta) := J(x,a'; \theta) \)
12: \textbf{end for}
13: \textbf{minimize} \( J(B, \theta) \) to update \( \theta \)
14: \textbf{minimize} \( J(B, \theta') \) to update \( \theta' \)
15: \textbf{save} the best models according to their performance on \( D_v \)
16: \textbf{until} no more matches
17: \textbf{until} both \( \theta \) and \( \theta' \) converge

of the considered node \( n_6 \), \textsc{TransX} exploits the context encoded by nodes \( n_1 \) and \( n_5 \) while \textsc{TransX}' makes use of the context of its sibling \( n_2 \). Intuitively, both nodes \( n_2 \) and \( n_3 \) are closely-related nodes of \( n_6 \), possessing important impact on the action prediction of \( n_6 \). Therefore, we believe that \textsc{TransX} and \textsc{TransX}' can complement each other.

**Experiments**

**Datasets**

We carry out experiments on the following datasets: 1) DJANGO (Oda et al. 2015). This dataset consists of 18,805 lines of Python source code, with each line paired with an NL utterance. We split the dataset into training/validation/test sets containing 16,000/1,000/1,805 instances, respectively. 2) ATIS. This dataset includes NL questions of a flight database, with each question is annotated with a lambda calculus query. Following previous studies (Yin and Neubig 2018, 2019; Xu et al. 2020), we use the standard splits of training/validation/test sets, which contains 4,473/491/448 instances, respectively. 3) GEO. It contains NL questions about US geography paired with corresponding Prolog database queries. We use the standard splits of 600/280 training/test instances. 4) IFTTT (Quirk, Mooney, and Galley 2015). It consists of if-this-then-that programs, paired with NL utterances of their purpose. The dataset is split into 68,083 training, 5,171 validation and 3,868 test instances.
Baseline Models

We apply mutual learning to simultaneously train Tranx and Tranx’, which are respectively referred to as ML-Tranx and ML-Tranx’, avoiding the description confusion. In addition to the conventional Tranx and Tranx’, and some commonly-used baselines (See Table 1), we compare the enhanced models with the following models:

- **Ensemble** (Sennrich et al. 2017). Using this method, we first employ depth-first and breadth-first model independently to obtain two k-best lists, and then re-score the combination of these two lists from these two models.
- **MTL-Tranx and MTL-Tranx’**. We apply multitask learning to jointly train Tranx and Tranx’ sharing the same encoder.
- **KD-Tranx**. We employ KD to transfer the knowledge of the Tranx’ with fixed parameters to enhance Tranx.
- **KD-Tranx’**. A Tranx’ enhanced by a fixed Tranx via KD.
- **ML2-Tranx**. Under our framework, two Tranx models with different initializations are jointly trained.
- **ML2-Tranx’**. Two Tranx’ models with different initializations are jointly trained under our framework.

Particularly, we report the average performance of two involved models for ML2-Tranx and ML2-Tranx’. We use the same experimental setup as (Yin and Neubig 2017). Specifically, we use 256 hidden units and 128-dimensional word vectors for NL utterance encoding, and tune the dimension of various embeddings on validation datasets for each corpus. We initialize all parameters by uniformly sampling within the interval [-0.1, 0.1]. Besides, we set the batch size as 10 and employ dropout after each layer, where the drop rate is sequentially set to 0.5, 0.3, 0.4, and 0.3 for our four datasets, respectively. To alleviate the instability of the model training, we run each model five times and report the average performance in terms of exact matching accuracy. Following the evaluation protocol in (Beltagy and Quirk 2016) for accuracies, we evaluate model performance on IFTTT at both channel and full parse tree (channel + function) levels.

Effect of $\lambda$

As shown in Equations 6 and 7, the coefficient $\lambda$ is an important hyper-parameter that controls relative impacts of the MLE loss and the divergence losses. Thus, we first investigate the effect of $\lambda$ on our proposed framework. To this end, we gradually vary $\lambda$ from 0 to 1 with an increment of 0.25 at each step, and report the performance of our models using different $\lambda$s on the validation datasets. Since there exists no validation set in GEO, we temporarily split the training data into two parts: 480 instances for training and 120 instances for validation, and use them to determine the optimal $\lambda$. However, with the optimal $\lambda$, we still use the whole training data in subsequent experiments.

Figure 3 shows the experimental results. According to the average performance of our models on four validation sets, we set $\lambda$ as 0.75, 0.5, 0.25 and 0.25 for our four datasets in all experiments thereafter.

Main Results

Table 1 reports the overall experimental results. Both models involved are improved under our framework. Besides, we can draw the following conclusions:

1. Our implemented Tranx achieves comparable performance to (Yin and Neubig 2019). Thus, we confirm that our reimplemented baseline is competitive.
2. Both ML-Tranx and ML-Tranx’ significantly surpass their respective important baselines: Tranx and Tranx’. Besides, both of our enhanced models surpass DRNN. These results strongly validate the effectiveness of our proposed training framework.
3. Both ML-Tranx and ML-Tranx’ outperforms Ensemble. This result is reasonable since Tranx and Tranx’ can interact with each other via mutual learning while they are independently trained when using ensemble.
4. Compared with multitask learning based models: MTL-Tranx and MTL-Tranx’, both ML-Tranx and ML-Tranx’ achieve better performance. This is due to the fact that in addition to sharing encoder parameters, mutual learning can refine model training by working as regularization terms, which has also been mentioned in (Romero et al. 2015; Clark et al. 2019).
5. ML-Tranx and ML-Tranx’ respectively show better performance than KD-Tranx and KD-Tranx’, verifying the necessity of multiple iterations of knowledge transfer.
6. ML-Tranx and ML-Tranx’ perform better in different datasets. The underlying reason is that each instance contains only three layers in IFTTT, where the effectiveness of pre-order decoding is greatly restricted.
formed well.

Table 2: Accuracy on different groups of DJANGO. Please
datasets into different groups according to their AST sizes,

\[ \text{RANX} \] 

<table>
<thead>
<tr>
<th>Model</th>
<th>DJANGO</th>
<th>ATIS</th>
<th>GEO</th>
<th>IFTTT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>Acc.</td>
<td>Acc.</td>
<td>Acc. CHANNEL / FULL TREE</td>
</tr>
<tr>
<td>DRNN (Alvarez-Melis and Jaakkola 2017)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>90.1 / 78.2</td>
</tr>
<tr>
<td>ASN (Rabinovich, Stern, and Klein 2017)</td>
<td>75.8</td>
<td>–</td>
<td>87.1</td>
<td>– / –</td>
</tr>
<tr>
<td>Seq2AST-YN17 (Yin and Neubig 2017)</td>
<td>74.1</td>
<td>87.7</td>
<td>88.2</td>
<td>– / –</td>
</tr>
<tr>
<td>Coarse2Fine (Dong and Lapata 2018)</td>
<td>77.3±0.4</td>
<td>87.6±0.1</td>
<td>88.6±1.0</td>
<td>– / –</td>
</tr>
<tr>
<td>TreeConv (Sun et al. 2019)</td>
<td>–</td>
<td>85.0</td>
<td>–</td>
<td>– / –</td>
</tr>
<tr>
<td>TreeGen (Sun et al. 2020)</td>
<td>–</td>
<td>89.1</td>
<td>89.6</td>
<td>– / –</td>
</tr>
<tr>
<td>Ensemble (Sennrich et al. 2017)</td>
<td>78.6±0.4</td>
<td>88.3±0.3</td>
<td>88.8±0.7</td>
<td>91.1±0.1 / 84.5±0.5</td>
</tr>
</tbody>
</table>

Table 1: Main experimental results. All results shown in the upper part are directly cited from their corresponding papers.

<table>
<thead>
<tr>
<th>Model</th>
<th>ATIS</th>
<th>GEO</th>
<th>IFTTT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[1, 10]</td>
<td>[11, 20]</td>
<td>[21, 30]</td>
</tr>
<tr>
<td>Ensemble</td>
<td>92.9</td>
<td>79.7</td>
<td>42.9</td>
</tr>
<tr>
<td>TRANX</td>
<td>93.0</td>
<td>76.7</td>
<td>45.7</td>
</tr>
<tr>
<td>MTL-TRANX</td>
<td>94.7</td>
<td>81.5</td>
<td>44.0</td>
</tr>
<tr>
<td>KD-TRANX</td>
<td>93.8</td>
<td>79.1</td>
<td>44.0</td>
</tr>
<tr>
<td>ML2-TRANX</td>
<td>93.8</td>
<td>80.9</td>
<td>44.3</td>
</tr>
<tr>
<td>ML-TRANX</td>
<td>95.0</td>
<td>82.2</td>
<td>48.6</td>
</tr>
<tr>
<td>TRANX’</td>
<td>92.4</td>
<td>79.1</td>
<td>42.3</td>
</tr>
<tr>
<td>MTL-TRANX’</td>
<td>94.8</td>
<td>81.1</td>
<td>44.6</td>
</tr>
<tr>
<td>KD-TRANX’</td>
<td>93.8</td>
<td>80.9</td>
<td>43.4</td>
</tr>
<tr>
<td>ML2-TRANX’</td>
<td>93.8</td>
<td>80.0</td>
<td>46.0</td>
</tr>
<tr>
<td>ML-TRANX’</td>
<td>94.0</td>
<td>81.8</td>
<td>43.4</td>
</tr>
</tbody>
</table>

Table 2: Accuracy on different groups of DJANGO. Please note that on the group [41, ∞), none of the models performed well.

(7) Compared with ML2-TRANX, our TRANX still exhibits better performance. Likewise, TRANX performs better than ML2-TRANX’. Hence, we confirm that TRANX and TRANX’ are indeed able to benefit each other.

Performance by the Size of AST

Further, in order to inspect the generality of our proposed framework, we follow Yin and Neubig (2017) to split datasets into different groups according to their AST sizes, and then compare the model performance at each group.

Tables 2, 3 and 4 display the results on DJANGO, ATIS and GEO, respectively.\(^1\)

\(^1\)We do not display the results on IFTTT, since all of its instances have the same AST size.
baselines respectively on almost all groups of datasets. Thus, we confirm again the effectiveness and generality of our proposed framework.

### Case Study

We compare the 1-best codes produced by different code generation models, so as to better understand how our model outperforms others. Table 5 shows an example of codes produced from Django dataset. We find that the model based on depth-first traversal tends to generate the wrong function names, while $\text{TRAX}'$ prefers to makes a mistake on the generation of argument. Although using various strategies, most of models are unable to generate completely correct code. By contrast, our model $\text{ML-TRAX}$ absorbs advantages of models with decoders based on different traversals and succeed in overcoming all these problems to generate more accurate code.

### Related Work

With the rapid development of deep learning, neural network based models have now become dominant in code generation. In this aspect, Ling et al., (2016) considered code generation as a conditional text generation task, which can be solved by a neural sequence-to-sequence model. To exploit syntactic and semantic constraints of the programs, both of which are important for code generation, more researchers resorted to neural Seq2Tree models transforming each NL utterance into a sequence of AST based grammar actions. For example, Yin and Neubig (2017; 2018) proposed Seq2Tree models to generate tree-structured MRCs using a series of tree-construction actions. Then, Sun et al. (2019; 2020) introduced CNN network and Transformer architecture to handle the long dependency problem.

From a different perspective, Yin and Neubig (2019) explored reranking an $N$-best list of candidate results for adequacy and coherence of the programs. Meanwhile, other approaches have been extensively investigated. Wei et al., (2019a) applied dual training to jointly model code summarization and code generation, where the specific intuitive correlation between these two tasks can be fully leveraged to benefit each other.

Significantly different from the above work, we further explore the Seq2Tree model with breadth-first traversal based decoding. More importantly, we introduce mutual learning based model training framework to iteratively enhance multiple models via knowledge distillation. Mutual learning has been successfully used in image recognition (Zhang et al. 2018b; Furlanello et al. 2018; Lan, Zhu, and Gong 2018), machine translation (Zeng et al. 2019), machine reading comprehension (Liu et al. 2020), and sentiment analysis (Xue, Zhang, and Zha 2020). However, to the best of our knowledge, our work is the first attempt to explore this approach for code generation. Among all models for code generation, the most related to our is (Alvarez-Melis and Jaakkola 2017). This work proposed to use doubly-recurrent decoders (vertical and horizontal LSTMs) trained on ASTs, of which outputs can be concatenated to exploit context in different dimensions. Compared with this work, ours still significantly differs from it in following aspects: 1) Our approach only affects the model training while has no impact on the model testing; 2) During the model testing, David Alvarez-Melis and Jaakkola. (2017) simultaneously use two LSTMs with more parameters while we only use one LSTM decoder, and 3) Experimental results show both of our enhanced models are able to surpass (Alvarez-Melis and Jaakkola 2017).

### Conclusion and Future Work

In this paper, we have explored the Seq2Tree model with breadth-first traversal based decoding and proposed a mutual learning based model training framework for code generation, where the models with different traversal based decodings can be enhanced simultaneously. Experimental results and in-depth analyses on several commonly-used datasets strongly demonstrate the effectiveness and generality of our proposed framework.

In the future, we plan to refine our framework via self-distillation (Wei et al. 2019b). Moreover, we will explore asynchronous bidirectional decoding (Zhang et al. 2018a; Su et al. 2019) to combine advantages of code generation models with different traversals based decoding (preorder traversal vs breadth-first traversal).

### Acknowledgements

The project was supported by National Natural Science Foundation of China (No. 62036004, No. 61672440), Natural Science Foundation of Fujian Province of China (No.2020J06001) and College-level Undergraduate Innovation Training Project of School of Informatics Xiamen University (No.2020Y1071).