Learning To Represent Edits

Pengcheng Yin*[1], Graham Neubig[1], Miltiadis Allamanis[2], Marc Brockschmidt[2], Alexander L. Gaunt[2]

[1]Language Technologies Institute, Carnegie Mellon University, USA [2]Microsoft Research Cambridge, UK

*Work primarily done as an intern at Microsoft Research

Learning to Represent and Apply Edits

Background Edits are ubiquitous in both source code and language

Before: `myDict[key] = obj.Value`
After: `myDict.Add(key, obj.Value)`

source code edits (e.g., GitHub commits)

Before: we have a great day
After: we had a great day

natural language edits

Research Questions

Edit Representation: How to learn edit patterns from data?

Neural Editor: How to apply an edit to a new input?

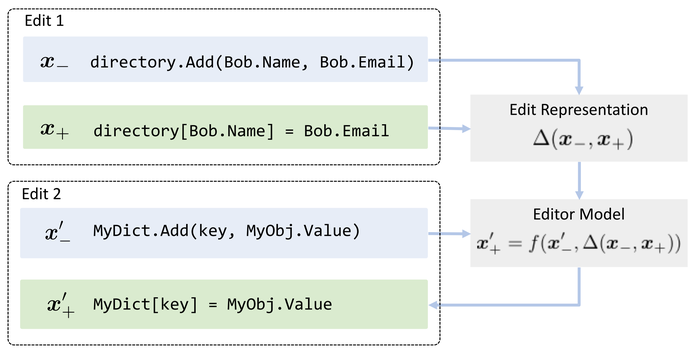
Contributions

Formalize the problem of data-driven learning of edits.

Present neural models that learn to represent and apply edits.

Release a large-scale dataset of code edits for future research.

Key Idea – End-to-End Learning to Edit



Learning Objective maximize the likelihood of generating the target x'_+ given the input x'_- and the "ground-truth" edit $\Delta(x'_-, x'_+)$

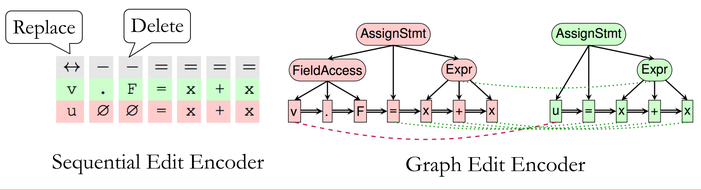
Edit Representations

Sequential Encoding of Edits

- Use deterministic diffing algorithm to get alignments of tokens in the source and target, which are encoded using bi-LSTMs

Graph Encoding of Edits

- Source and target data is represented as syntax trees
- Add alignment edges between edited nodes. Encode the graph using graph neural networks (Li et al., 2016)

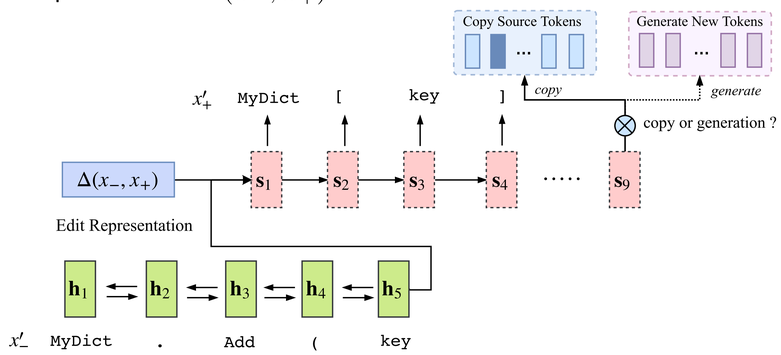


Neural Editors

Overview Given input x'_-, and edit representation $\Delta(x_-, x_+)$, a neural editor applies the edit to x'_- and generates the updated input x'_+

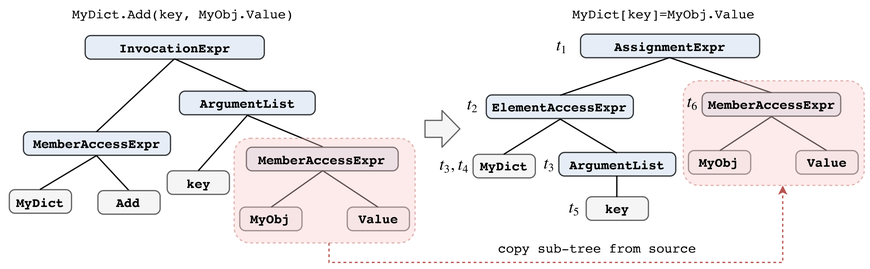
Sequence-to-Sequence Neural Editors

- Seq2Seq editors encode the input data as a sequence of tokens
- A recurrent neural network generates the target x'_- using the input x'_+ and the edit representation $\Delta(x_-, x_+)$

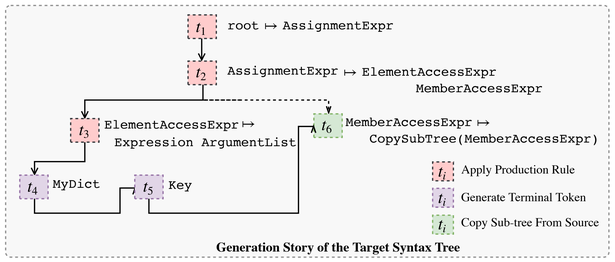


Graph-to-Tree Neural Editors

Motivation the data to edit (e.g., source code) usually has strong underlying structure. How to utilize the structural information to better predict the edited output?



Graph2Tree editors transduce the tree-structured input to the target tree via a sequence of tree-constructing actions (Yin et al., 2018)



- Source and target data (code) is represented as structured abstract syntax trees
- Source data is encoded using gated graph neural networks (Allamanis et al., 2018)

GitHubEdits Code Edit Dataset

- 110K pairs of C# code edits collected from GitHub commit histories
- Each edit involves at most three consecutive lines of code
- Shipped with parsed abstract syntax trees for each sample

Scan me

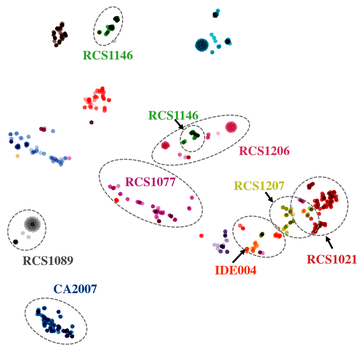
Experiments

Quality of Edit Representations

1. t-SNE Visualization Systems trained on the noisy GitHubEdits, and tested on 2.8K edit pairs with known, labeled edit categories (16 in total). Examples:

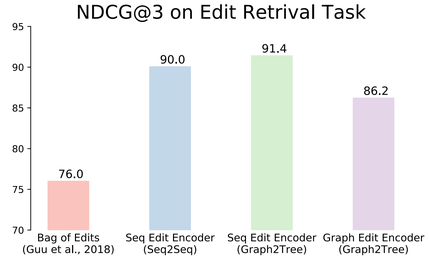
Before `x != null && x.StartsWith("a")`
After `x?.StartsWith("a")`
Category Use conditional Access (RCS1146)

Before `i = i + 1`
After `i++`
Category Use --/++ operator (RCS1089)



2. Edit Retrieval Task Given a seed edit representation, retrieve and evaluate the relevance of retrieved neighboring edits

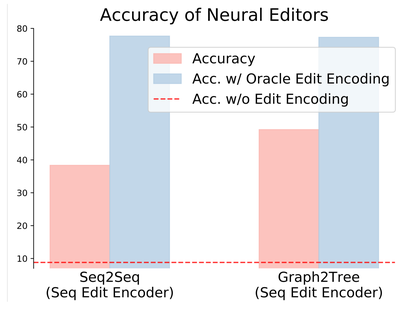
Manually annotate the relevance with a scale of 3 (*highly relevant, relevant, irrelevant*)



Precision of Neural Editors

End-to-End Evaluation Given a learned edit encoding $\Delta(x_-, x_+)$, apply the edit to a similar input x'_- to generate the edited input x'_+

Compare with the upper-bound accuracy of using the oracle edit encoding $\Delta(x'_-, x'_+)$



Clustering Edits on GitHub Commits and Wikipedia Edit History

Desc. Optimize LINQ Queries
Before `V0.Customers.Where(V1 => V1.CustomerID == LITERAL).FirstOrDefault()`
After `V0.Customers.FirstOrDefault(V1 =>V1.CustomerID == LITERAL)`

Before `this.V0.Where(V1 => V1.CanDeserialize(V2)).FirstOrDefault()`
After `this.V0.FirstOrDefault(V1 => V1.CanDeserialize(V2))`

Before `V0.TypeConverters.Where(V1 => V1.CanConvertTo(V2,V0)).FirstOrDefault()`
After `V0.TypeConverters.FirstOrDefault(V1 => V1.CanConvertTo(V2, V0))`

Desc. Switch from `Assert.Equal` to `Assert.Empty`
Before `Assert.Equal(0, V0.ProjectIds.Count)`
After `Assert.Empty(V0.ProjectIds)`

Before `Assert.Equal(0, V0.ProjectReferences.Count())`
After `Assert.Empty( V0.ProjectReferences)`

Before `Assert.Equal(0, V0.Messages.Count)`
After `Assert.Empty(V0.Messages)`

Sampled edit clusters from GitHubEdits (omitting variable assignments for clarity)

Desc. Add a parenthetical expression *also ... as*

mid-state regional airport, +also known as mid-state airport,+ is a small airport on in rush township…

islamic culture, +also known as saracenic culture,+ is a term primarily used in secular academia…

birds of prey, +also known as raptors,+ are birds that hunt for food primarily via flight…

Desc. Add a Person's Middle Name

isaiah +marcus+ rankin (born 22 may 1978 in london) is an english professional footballer

audrey +Kathleen+ brown (born 24 may , 1913) is a british athlete who competed mainly in the 100 metres .

monique +edith+ lamoureux (born july 3 , 1989) is an american ice hockey player .

Sampled edit clusters (+insertion+ edits) from WikiAtomicEdits dataset (Faruqui et al., 2018)