Fine-grained and Accurate Source Code Differencing

Problem Statement

- Existing approaches usually represent code changes or edit operations as add line or delete line actions
- Such representations are not precise
  - E.g., code move or update is not properly represented

Contributions

- GumTree—a novel efficient AST differencing algorithm that includes move actions
- An automated evaluation of GumTree
- A manual evaluation to compare GumTree vs. textual diff
- An automated evaluation to compare GumTree vs. ?

The GumTree Algorithm

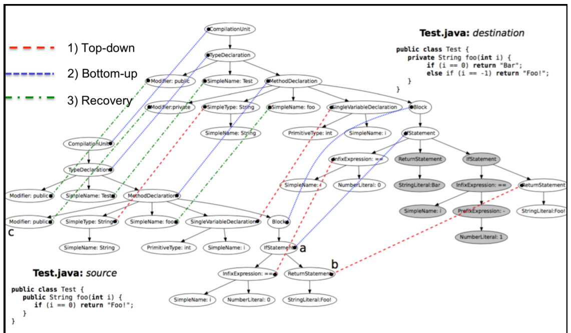
- 1. A greedy top-down algorithm to find isomorphic sub-trees of decreasing height. Mappings are established between the nodes of these isomorphic subtrees. They are called anchors mappings.

The GumTree Algorithm (cont'd)

- 2. A bottom-up algorithm where two nodes match (called a container mapping) if their descendants (children of the nodes, and their children, and so on) include a large number of common anchors. When two nodes match, we finally apply an optimal algorithm to search for additional mappings (called recovery mappings) among their descendants.

The GumTree Algorithm (cont'd)

- 3. Recovery Mappings: to find additional mappings between leaf nodes and similar nodes
- 4. Generate edit operations for the unmatched nodes:
  - Insert
  - Delete
  - Update
  - Move



Test.java: *destination*

```
public class Test {
   private String foo(int i) {
      if (i == 0) return "Bar";
      else if (i == -1) return "Foo!";
   }
}
```

Test.java: *source*

```
public class Test {
   public String foo(int i) {
      if (i == 0) return "Foo!";
   }
}
```

7

Top-Down Phase

- Start with the roots and check if they are isomorphic or identical. If not, the children nodes are then tested
- To identify the unchanged part
- Implementation
  - By hardcoding subtrees, the isomorphism test's complexity is O(1)
  - The worst-case complexity is O(n^2)

8

Bottom-Up Phase

- Search for container mappings, that are established when two nodes have a significant number of matching descendants

$$\mathrm{dice}(t_1, t_2, \mathcal{M}) = \frac{2 \times |\{t_1 \in s(t_1) | (t_1, t_2) \in \mathcal{M}\}}{|s(t_1)| + |s(t_2)|}$$

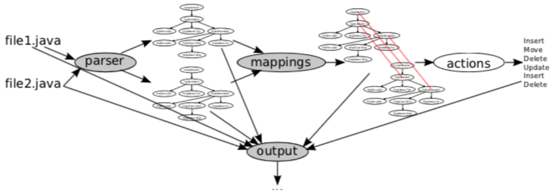
9

Recovery Mappings

- Given two trees, find their additional mappings between the descendants,
  - remove the matched descendants, and
  - apply an optimized algorithm to find a shortest edit script without move actions

10

Architecture



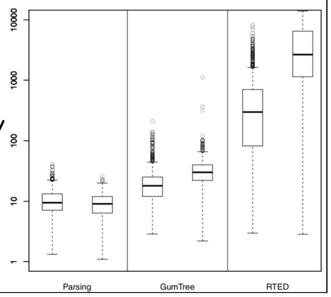
- Parser: Java, JavaScript, R, and C
- Mappings: GumTree, ChangeDistiller, XYDiff, RTED
- Output: XML representation of AST, web-based view of an edit script, XML representation of an edit script

11

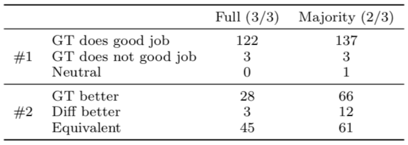
Evaluation

Comparison between GumTree, textual diff, and RTED

- The median of parsing time is 10
- Computing an edit script is only slightly slower than just parsing the files (median at 18 for Jenkins and 30 for JQuery



Manual Evaluation



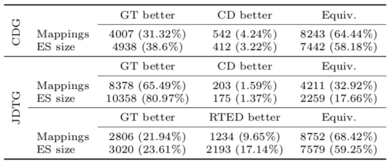
		Full (3/3)	Majority (2/3)
#1	GT does good job	122	137
	GT does not good job	3	3
	Neutral	0	1
#2	GT better	28	66
	Diff better	3	12
	Equivalent	45	61

Table 1: Agreements of the manual inspection of the 144 transactions by three raters for Question #1 (top) and Question #2 (bottom).

- GumTree's output is sometimes better than textual diff

Automatic Evaluation

- More matches = better

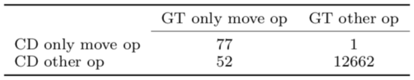


		GT better	CD better	Equiv.
CDG	Mappings	4007 (31.32%)	542 (4.24%)	8243 (64.44%)
	ES size	4938 (38.6%)	412 (3.22%)	7442 (58.18%)
JDTG		GT better	CD better	Equiv.
	Mappings	8378 (65.49%)	203 (1.59%)	4211 (32.92%)
	ES size	10358 (80.97%)	175 (1.37%)	2259 (17.66%)
		GT better	RTED better	Equiv.
	Mappings	2806 (21.94%)	1234 (9.65%)	8752 (68.42%)
	ES size	3020 (23.61%)	2193 (17.14%)	7579 (59.25%)

Table 2: Number of cases where GumTree is better (resp. worse and equivalent) than ChangeDistiller (top, middle) and RTED (bottom) for 2 metrics, number of mappings and edit script size (ES size), at the CDG granularity (top) and JDTG granularity (middle, bottom).

Automatic Evaluation (cont'd)

- GumTree generates smaller edit scripts in most cases than RTED and ChangeDistiller
  - 130 elements include move-only actions



	GT only move op	GT other op
CD only move op	77	1
CD other op	52	12662

Table 3: Comparison of the number of move operations from GumTree and ChangeDistiller for 12 792 file pairs to be compared.