An evaluation of Diagnosis in a Learning Environment for Object-Oriented Modeling

Ludovic AUXEPAULES, Dominique PY

LIUM - Laboratoire d'Informatique de l'Université du Maine (France)

ICALT 2010 – July 7th 2010
Context of this work

- Project of the LIUM laboratory: « Interaction and knowledge »

- **Participants:** Dominique Py, Mathilde Alonso, Thierry Lemeunier and Ludovic Auxepaules

- **Goal of the project:** designing models, methods and tools for object-oriented modeling learning environments

- **Application:** the *Diagram* environment
Outline

- The *Diagram* environment

- The ACDC matching method
  (*Automatic Class Diagrams Comparator*)

- A diagnosis example produced by ACDC

- An ACDC evaluation with *Diagram* used on ecological context

- Conclusions and perspectives
The **Diagram** environment

- An open UML class diagrams editor
  - Allowing the student to work with the problem text and the diagram together
  - Providing specific interaction modes and help features for novice users

- A three-step method for solving modeling problems

![Diagram environment diagram](image)
A pen and a felt-pen are two concepts with common attributes: color, brand name, etc. A felt-pen has a top. Both pen and felt-pen have a body with some properties. Pens and felt-pens are used by a person and belong to a person. There is a specific felt-pen that is an eraser felt-pen.

Example of feedback messages in **Diagram**

- **Incorrect Items**
  - Message #1
    - Notification
    - You say that Felt-pen is a type of Eraser_felt-pen
    - Question
    - Is Felt-pen a type of Eraser_felt-pen?
    - Proposition
    - I would rather say that Eraser_felt-pen is a type of Felt-pen

- **Feedback messages**

- **Omitted Items**
  - Message #4
    - Notification
    - The diagram is incomplete: a main concept is missing
    - Question
    - Did you represent all the concepts in the diagram?

**Diagram environment** | **ACDC matching method** | **Diagnosis example** | **ACDC evaluation** | **Conclusions and perspectives**
The proposition of diagnosis within *Diagram*

- **Modeling and *Diagram* contexts**
  - Diagnosis in our context: the system ability to analyse student's answers
  - No pedagogical solver in open-ended domains like modeling
  - Calculus time should be *fast enough* for synchronous pedagogical feedbacks

- **A diagnosis based on a models matching method**
  - Comparison of the *diagram built by the learner* with a *reference model* supplied by an expert
  - Production of a *differences list* between the models (no errors)
ACDC matching method
(Automatic Class Diagrams Comparator)

- An **automatic** and **customizable** matching method [Auxepaules 09]

![Diagram](image)

- An hybrid matcher that combines
  - String-based similarity of namespaces and type similarity constraints
  - Element-level and structure-level matching techniques [Shvaiko & Euzenat 05]

- Univalent or multivalent mappings at all model granularity levels
  - It fully or partially matches **one or more** structures of a model to **one or more** structures of another model and mutually
3 sequential steps of ACDC matching method

1. Schematization of inputs models into structural patterns (simple or complex)
   - Trees and graphs algorithms: search, cover and sort of roots, leafs, ways...

2. Evaluation of local similarities and differences of each patterns couple per type
   - Similarity function that combines lots of criteria: names, context, specific properties...

3. Choice of one mapping of models patterns and differences
   - Greedy process without backtracking
Our differences taxonomy [Auxepaules 09]

- **Univalent** difference: partial match of two single patterns (1:1 matching cardinality)
- **Multivalent** difference: partial match of a patterns group (n:m matching cardinality)

Those differences are converted into pedagogical differences for elaborating feedback messages [Py et al. 08]
A diagnosis example produced by ACDC

- 12 full univalent matches (without structural differences)

- 4 main content and structural differences used by Diagram to produce pedagogical feedback messages
**Example of a compound difference**

**Differences identified by ACDC**

<table>
<thead>
<tr>
<th>Differences</th>
<th>Pedagogical feedbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>{have (body---felt pen) has (body---pen)} SPLIT</td>
<td>Duplication and transfer of a relationship</td>
</tr>
<tr>
<td>{has (Body---Pencil)}</td>
<td></td>
</tr>
<tr>
<td>{have (body---felt pen)} TRANSFER LOWER</td>
<td></td>
</tr>
<tr>
<td>{has (Body---Pencil)}</td>
<td></td>
</tr>
<tr>
<td>{has (body---pen)} TRANSFER LOWER</td>
<td></td>
</tr>
<tr>
<td>{has (Body---Pencil)}</td>
<td></td>
</tr>
</tbody>
</table>
Example of an other compound difference

**Student's diagram**

**Reference diagram**

<table>
<thead>
<tr>
<th>Differences identified by ACDC</th>
<th>Pedagogical feedbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>{felt pen::top} REPLACEMENT {Top}</td>
<td>Misrepresentation of a class and omission of linked elements</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>{felt pen::top} INCOMPATIBLE_NATURE {Top}</td>
<td></td>
</tr>
<tr>
<td>OMISSION {belongs to (Person---Pencil)}</td>
<td></td>
</tr>
</tbody>
</table>
Example of two simple differences

<table>
<thead>
<tr>
<th>Differences identified by ACDC</th>
<th>Pedagogical feedbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMISSION {belongs to (Person---Pencil)}</td>
<td>Omission of a relationship</td>
</tr>
<tr>
<td>{Pencil} VOID {Pencil}</td>
<td>Misrepresentation of a class</td>
</tr>
<tr>
<td>{Pencil} NOT_ABSTRACT_TO_ABSTRACT {Pencil}</td>
<td>(a class instead of an abstract class)</td>
</tr>
</tbody>
</table>

Diagram environment  | ACDC matching method  | Diagnosis example  | ACDC evaluation  | Conclusions and perspectives  |
---------------------|-----------------------|--------------------|------------------|-------------------------------|

Ludovic Auxepaules  13 / 20
1st evaluation of ACDC within *Diagram*

- **Previous evaluations**
  - Tests of usability, interaction and pedagogical functionalities of *Diagram* (without diagnosis system) [Alonso *et al*. 08]
  - Off-line evaluation of ACDC (not in *Diagram*) [Auxepaules *et al*. 2008]

- **Protocol of the evaluation of ACDC within *Diagram***
  - 18 novice students in 2nd year of University (DEUST)
  - 4 practice sessions of 3 hours of modeling in *Diagram*
  - Reference diagrams built by OOM teacher
  - Sessions, built diagrams, diagnosis calls and feedbacks have been recorded

- **Evaluation of feedback messages reported in** [Alonso & Py 09]
Match quality measures

- Comparison of the automatically identified matches (ACDC outputs) with the manually determined real matches
  - *False positives A*: matches needed but not identified by the matcher
  - *True positives B*: correct matches automatically predicted by the matcher
  - *False negatives C*: matches falsely proposed or mistyped by the matcher

- Measures of quality [Do & Rahm 07] [Giunchiglia et al. 07] [Melnik et al. 02]
  - *Precision* = \(|B| / (|B| + |C|)\) → an accuracy or fidelity measure
  - *Recall* = \(|B| / (|B| + |A|)\) → a completeness measure
  - *F-Measure* → an harmonic mean of *Precision* and *Recall*
  - *Overall* = \((|B| - |C|) / (|B| + |A|)\) → an evaluation of the post-match effort needed for adding missed matches and removing false ones

Comparison of the automatically identified matches (ACDC outputs) with the manually determined real matches

- False positives A: matches needed but not identified by the matcher
- True positives B: correct matches automatically predicted by the matcher
- False negatives C: matches falsely proposed or mistyped by the matcher

Measures of quality [Do & Rahm 07] [Giunchiglia et al. 07] [Melnik et al. 02]

- Precision = \(|B| / (|B| + |C|)\) → an accuracy or fidelity measure
- Recall = \(|B| / (|B| + |A|)\) → a completeness measure
- F-Measure → an harmonic mean of Precision and Recall
- Overall = \((|B| - |C|) / (|B| + |A|)\) → an evaluation of the post-match effort needed for adding missed matches and removing false ones
Results at the diagram level

<table>
<thead>
<tr>
<th>Number and % of diagnosis calls where</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality Measure Result = 1 (best)</td>
<td>103</td>
<td>60</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>71.6%</td>
<td>41.7%</td>
<td>38.2%</td>
<td>38.2%</td>
</tr>
<tr>
<td>0.85 ≤ Quality Measure Result &lt; 1</td>
<td>30</td>
<td>65</td>
<td>78</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>20.8%</td>
<td>45.1%</td>
<td>54.2%</td>
<td>41%</td>
</tr>
<tr>
<td>0.7 ≤ Quality Measure Result &lt; 0.85</td>
<td>11</td>
<td>19</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>7.6%</td>
<td>13.2%</td>
<td>6.9%</td>
<td>15.3%</td>
</tr>
<tr>
<td>Quality Measure Result &lt; 0.7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>0%</td>
<td>0%</td>
<td>0.7%</td>
<td>5.5%</td>
</tr>
</tbody>
</table>

- Good results for the 144 diagnosis calls (144 different student diagrams)
  - 38% of diagnosis calls outputs are perfect
  - More than 92% of diagnosis calls outputs are relevant at 85%
  - At least one mismatch on only 28% of 144 diagnosis calls
  - At least one omitted match on 58% of 144 diagnosis calls

- Diagnosis calculi time from 0.2s to 6s with an average of 2s
## Results by type of structural differences

<table>
<thead>
<tr>
<th>Differences</th>
<th>A: false negatives</th>
<th>B: true positives</th>
<th>C: false positives</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split</td>
<td>43</td>
<td>71</td>
<td>23</td>
<td>0.76</td>
<td>0.63</td>
</tr>
<tr>
<td>Merge</td>
<td>6</td>
<td>47</td>
<td>22</td>
<td>0.68</td>
<td>0.89</td>
</tr>
<tr>
<td>Void</td>
<td>31</td>
<td>3069</td>
<td>52</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>Replacement</td>
<td>62</td>
<td>416</td>
<td>33</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>Transfer</td>
<td>141</td>
<td>133</td>
<td>18</td>
<td>0.88</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>283</strong></td>
<td><strong>3736</strong></td>
<td><strong>148</strong></td>
<td><strong>0.96</strong></td>
<td><strong>0.93</strong></td>
</tr>
</tbody>
</table>

- Most of matches are **strict** (without difference of structure)

- Results related to the biggest structure alterations of models are well
  - 70% of **multivalent matches** are identified and 72% are correct
Results with regard to number of differences

- Precision and Recall trends of ACDC decrease according to the number of differences between compared diagrams.
- Precision results are better than Recall Results.
  - A wrong match can replace one or more correct matches.

<table>
<thead>
<tr>
<th>Diagram environment</th>
<th>ACDC matching method</th>
<th>Diagnosis example</th>
<th>ACDC evaluation</th>
<th>Conclusions and perspectives</th>
</tr>
</thead>
</table>

Quality measure value

Number of identified differences
Conclusions and perspectives

- An evaluation overview shows that ACDC worked fairly well but requires to be improved in some specific situations.

- Use of several reference diagrams to analyze the student's diagram:
  - Complete / part of diagrams and correct / erroneous diagrams
  - ACDC can already compare more than two models and choose one reference at the end of the 2nd step of similarities and differences patterns evaluation.

- Extend to others kinds of model or domains:
  - Direct transposition for Entity-Relationship models of Data-Bases.

- Use of ACDC system outputs in other context: teacher needs.
Thank you for your attention