Evaluating Storage and Reasoning Systems

Mikalai Yatskevich
Information Systems Group
Oxford University Computing Laboratory, United Kingdom
mikalai.yatskevich@comlab.oc.ac.uk
Index

• Evaluation scenarios
• Evaluation descriptions
• Test data
• Tools
• Results
• Conclusion
Advanced reasoning system

• Description logic based system (DLBS)
• Standard reasoning services
  – Classification
  – Class satisfiability
  – Ontology satisfiability
  – Logical entailment
Existing evaluations

• Datasets
  – Synthetic generation
  – Hand crafted ontologies
  – Real-world ontologies

• Evaluations
  – KRSS benchmark
  – TANCS benchmark
  – Gardiner dataset
Evaluation criteria

- **Interoperability**
  - the capability of the software product to interact with one or more specified systems
  - a system must
    - conform to the standard input formats
    - be able to perform standard inference services

- **Performance**
  - the capability of the software to provide appropriate performance, relative to the amount of resources used, under stated conditions
Evaluation metrics

- Interoperability
  - Number of tests passed without parsing errors
  - Number of inference tests passed
- Performance
  - Loading time
  - Inference time
Class satisfiability evaluation

- Standard inference service that is widely used in ontology engineering
- The goal: to assess both DLBS’s interoperability and performance
- Input
  - OWL ontology
  - One or several class IRIs
- Output
  - TRUE the evaluation outcome coincide with expected result
  - FALSE the evaluation outcome differ from expected outcome
  - ERROR indicates IO error
  - UNKNOWN indicates that the system is unable to compute inference in the given timeframe
Class satisfiability evaluation

DLBS

Evaluation module

OWL 2 ontology

Load ontology

Measure time

T1:ns

Class URIs

Class satisfiability

Measure time

T2:ns

Measure time

Tn:ns

Is the last test?

TRUE/FALSE/ERROR/UNKNOWN

No

Yes

Save evaluation result
Ontology satisfiability evaluation

• Standard inference service typically carried out before performing any other reasoning task
• The goal: to assess both DLBS’s interoperability and performance
• Input
  – OWL ontology
• Output
  – TRUE the evaluation outcome coincide with expected result
  – FALSE the evaluation outcome differ from expected outcome
  – ERROR indicates IO error
  – UNKNOWN indicates that the system is unable to compute inference in the given timeframe
Ontology satisfiability evaluation
Classification evaluation

- Inference service that is typically carried out after testing ontology satisfiability and prior to performing any other reasoning task.
- The goal: to assess both DLBS’s interoperability and performance.
- Input:
  - OWL ontology
- Output:
  - OWL ontology
  - ERROR indicates IO error
  - UNKNOWN indicates that the system is unable to compute inference in the given timeframe.
Logical entailment evaluation

- Standard inference service that is the basis for query answering
- The goal: to assess both DLBS’s interoperability and performance
- Input
  - 2 OWL ontologies
- Output
  - TRUE the evaluation outcome coincide with expected result
  - FALSE the evaluation outcome differ from expected outcome
  - ERROR indicates IO error
  - UNKNOWN indicates that the system is unable to compute inference in the given timeframe
Logical entailment

DLBS

- Load ontology
- Logical entailment
- TRUE/FALSE/ERROR/UNKNOWN

Evaluation module

- Measure time
- Measure time
- Measure time
- Save evaluation result
Storage and reasoning systems evaluation component

• SRS component is intended to evaluate the description logic based systems (DLBS)
  – Implementing OWL-API 3 de-facto standard for DLBS
  – Implementing SRS SEALS DLBS interface
• SRS supports test data in all syntactic formats supported by OWL-API 3
• SRS saves the evaluation results and interpretations in MathML 3 format
DLBS interface

- Java methods to be implemented by system developers
  - OWLOntology loadOntology(IRI iri)
  - boolean isSatisfiable(OWLOntology onto, OWLClass class)
  - boolean isSatisfiable(OWLOntology onto)
  - OWLOntology classifyOntology(OWLOntology onto)
  - URI saveOntology(OWLOntology onto, IRI iri)
  - boolean entails(OWLOntology onto1, OWLOntology onto2)
Testing Data

• The ontologies from the Gardiner evaluation suite.
  – Over 300 ontologies of varying expressivity and size.
• Various versions of the GALEN ontology
• Various ontologies that have been created in EU funded projects, such as SEMINTEC, VICODI and AEO
• 155 entailment tests from OWL 2 test cases repository
Evaluation setup

- 3 DLBSs
  - FaCT++ C++ implementation of FaCT OWL DL reasoner
  - HermiT Java based OWL DL reasoner utilizing novel hypertableau algorithms
  - Jcel Java based OWL 2 EL reasoner

- 2 AMD Athlon(tm) 64 X2 Dual Core Processor 4600+ machines with 2GB of main memory
  - DLBSs were allowed to allocate up to 1 GB
## Evaluation results: Classification

<table>
<thead>
<tr>
<th></th>
<th>FaCT++</th>
<th>HermiT</th>
<th>jcel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALT, ms</td>
<td>68</td>
<td></td>
<td>856</td>
</tr>
<tr>
<td>ART, ms</td>
<td>15320</td>
<td></td>
<td>2144</td>
</tr>
<tr>
<td>TRUE</td>
<td>160</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>FALSE</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>ERROR</td>
<td>47</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>3</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>FaCT++</td>
<td>HermiT</td>
<td>jcel</td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>ALT, ms</td>
<td>1047</td>
<td>255</td>
<td>438</td>
</tr>
<tr>
<td>ART, ms</td>
<td>21376</td>
<td>517043</td>
<td>1113</td>
</tr>
<tr>
<td>TRUE</td>
<td>157</td>
<td>145</td>
<td>15</td>
</tr>
<tr>
<td>FALSE</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ERROR</td>
<td>36</td>
<td>35</td>
<td>5</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>16</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>
## Evaluation results: Ontology satisfiability

<table>
<thead>
<tr>
<th></th>
<th>FaCT++</th>
<th>HermiT</th>
<th>jcel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALT, ms</td>
<td>1315</td>
<td></td>
<td>708</td>
</tr>
<tr>
<td>ART, ms</td>
<td>25175</td>
<td></td>
<td>1878</td>
</tr>
<tr>
<td>TRUE</td>
<td>134</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>FALSE</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>ERROR</td>
<td>45</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
## Evaluation results: Entailment

<table>
<thead>
<tr>
<th></th>
<th>FaCT++</th>
<th>HermiT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALT, ms</td>
<td>14</td>
<td>33</td>
</tr>
<tr>
<td>ART, ms</td>
<td>1</td>
<td>20673</td>
</tr>
<tr>
<td>TRUE</td>
<td>46</td>
<td>119</td>
</tr>
<tr>
<td>FALSE</td>
<td>67</td>
<td>14</td>
</tr>
<tr>
<td>ERROR</td>
<td>34</td>
<td>9</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
# Evaluation results: Non entailment

<table>
<thead>
<tr>
<th></th>
<th>FaCT++</th>
<th>HermiT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALT, ms</td>
<td>47</td>
<td>92</td>
</tr>
<tr>
<td>ART, ms</td>
<td>5</td>
<td>127936</td>
</tr>
<tr>
<td>TRUE</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>FALSE</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ERROR</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Conclusion

• Errors:
  – datatypes not supported in the systems
  – syntax related: a system was unable to register a role or a concept
  – expressivity errors

• Execution time is dominated by small number of hard problems