Improving UIMA Debug Capabilities

Table of Contents

1. Provenance Tracking of UIMA CAS Content

1.1 Use-cases
Users have requested the ability to track which Analysis Engines (AEs) created or modified the feature structures in a CAS. Such information would be useful for debugging as well as understanding the behavior of a complex assembly of component AEs.

1.2 Proposal
Provide ability to save changes to the CAS done by a component as an XMI "difference" format (see related proposal about "delta CAS" in Apache UIMA /Wiki). This information will be stored as files and/or viewed directly using a modified AnnotationViewer type GUI.

1.3 XMI based Specification of Journal xmiCAS
The XMI standard specifies a possibility for transmitting metadata differences by using Add, Replace, and Delete elements. The goal is to provide a mechanism for specifying the differences between documents in a way that the entire document does not need to be transmitted each time.

The following input text is used to illustrate some details of how provenance would work.

```
UIT Seminar: Challenges in Speech Recognition
          August 8, 2003   10:30 AM - 11:30 AM
          Lawrence Rabiner , Associate Director CAIP, Rutgers
          University, Professor Univ. of Santa Barbara
          Yorktown 20-04
          Availability: Open
```

1.3.1 Standard xmiCAS
The following is the original xmiCAS produced by MeetingFinderAggregate (example descriptor in UIMA-AS distribution, in $UIMA_HOME/examples/deploy /as/MeetingFinderAggregate.xml). For readability, the annotations produced by TokenAndSentence are omitted.

```
<?xml version="1.0" encoding="UTF-8"?>
<xmi:XMI xmlns:tutorial="http:///org/apache/uima/tutorial.ecore"
      xmlns:tcas="http:///uima/tcas.ecore"
      xmlns:tokenizer="http:///org/apache/uima/examples/tokenizer.ecore"
      xmlns:xmi="http://www.omg.org/XMI"
      xmlns:cas="http:///uima/cas.ecore" xmi:version="2.0">
      <cas:NULL xmi:id="0"/>
      <cas:Sofa xmi:id="1" sofaNum="1" sofaID="_InitialView" mimeType="text" sofaString="UIT Seminar:
Challenges in Speech Recognition&#10;  August 8, 2003   10:30 AM - 11:30 AM &#10;  Lawrence Rabiner , Associate
Director CAIP, Rutgers &#10;  University, Professor Univ. of Santa Barbara   &#10;  Yorktown 20-043&#10;
Availability: Open &#10;"/>
      <tcas:DocumentAnnotation xmi:id="8" sofa="1" begin="0" end="231" language="en"/>
      <tutorial:RoomNumber xmi:id="13" sofa="1" begin="202" end="208" building="Watson - Yorktown"/>
      <tutorial:TimeAnnot xmi:id="18" sofa="1" begin="65" end="74" shortDateString="1/1/70"/>
      <tutorial:TimeAnnot xmi:id="23" sofa="1" begin="76" end="85" shortDateString="1/1/70"/>
      <tutorial:DateAnnot xmi:id="28" sofa="1" begin="48" end="62" shortDateString="8/8/03"/>
      <tutorial:Meeting xmi:id="33" sofa="1" begin="48" end="208" room="13" date="28" startTime="18" endTime="
23"/>
      <cas:View sofa="1" members="8 13 18 23 28 33"/>
</xmi:XMI>
```

1.3.2 Sample of Journal xmiCAS

The following is the proposed journal xmiCAS. For readability, this is a simplified format that will be different from the actual implementation.

```
<?xml version="1.0" encoding="UTF-8"?>
<xmi:XMI xmlns:tutorial="http:///org/apache/uima/tutorial.ecore"
    xmlns:tcas="http:///uima/tcas.ecore"
    xmlns:tokenizer="http:///org/apache/uima/examples/tokenizer.ecore"
    xmlns:xmi="http://www.omg.org/XMI"
    xmlns:cas="http:///uima/cas.ecore" xmi:version="2.0">
    xmlns:deltaCas="http:///uima/deltaCas.ecore" xmi:version="2.0">
    <cas:NULL xmi:id="0"/>

    <deltaCas:callFlowController/>  <!-- Flow controller runs before each delegate. -->

    <deltaCas:callAE key="CollectionReader" path="MeetingFinderAggregate" >
        <xmi:Add addition="1" />
        <cas:Sofa xmi:id="1" sofaNum="1" sofaID="_InitialView" mimeType="text" sofaString="UIT Seminar:
Challenges in Speech Recognition&#10;  August 8, 2003   10:30 AM - 11:30 AM &#10;  Lawrence Rabiner , Associate
Director CAIP, Rutgers &#10;  University, Professor Univ. of Santa Barbara   &#10;  Yorktown 20-043&#10;
Availability: Open &#10;"/>

        <xmi:Add addition="8" />
        <tcas:DocumentAnnotation xmi:id="8" sofa="1" begin="0" end="231" language="en"/>

        <xmi:Add addition="_InitialView" />
        <cas:View  xmi:id="_InitialView" sofa="1" members="8"/>
    </deltaCas:callAE>

    <deltaCas:callFlowController/>

    <deltaCas:callAE key="RoomNumber"  path="MeetingFinderAggregate/MeetingDetectorTae">
        <xmi:Add addition="13" />
        <tutorial:RoomNumber xmi:id="13" sofa="1" begin="202" end="208" building="Watson - Yorktown"/>

        <xmi:Replace replacement="_InitialView" >
            <target href="_InitialView"  />
        </xmi:Replace>
        <cas:View xmi:id="_InitialView" sofa="1" members="8 13"/>
    </deltaCas:callAE>

    <deltaCas:callFlowController/>

    <deltaCas:callAE key="DateTime" path="MeetingFinderAggregate/MeetingDetectorTae" >
        <xmi:Add addition="18"  />
        <tutorial:TimeAnnot xmi:id="18" sofa="1" begin="65" end="74" shortDateString="1/1/70"/>

        <xmi:Add addition="23"  />
        <tutorial:TimeAnnot xmi:id="23" sofa="1" begin="76" end="85" shortDateString="1/1/70"/>

        <xmi:Add addition="28" />
        <tutorial:DateAnnot xmi:id="28" sofa="1" begin="48" end="62" shortDateString="8/8/03"/>

        <xmi:Replace replacement="_InitialView" >
            <target href="_InitialView"  />
        </xmi:Replace>
        <cas:View xmi:id="_InitialView" sofa="1" members="8 13 18 23 28"/>
    </deltaCas:callAE>

    <deltaCas:callFlowController/>

    <deltaCas:callAE key="Meeting" path="MeetingFinderAggregate/MeetingDetectorTae" >
        <xmi:Add addition="33"  />
        <tutorial:Meeting xmi:id="33" sofa="1" begin="48" end="208" room="13" date="28" startTime="18" endTime="
23"/>

        <xmi:Replace replacement="_InitialView" >
            <target href="_InitialView"  />
        </xmi:Replace>
        <cas:View xmi:id="_InitialView" sofa="1" members="8 13 18 23 28 33"/>
    </deltaCas:callAE>
</xmi:XMI>
```

1.4 Open Issues

1.4.1 xmi:id
From the XMI specification, the ID (the value of the xmi:id) of each feature structure is required to be unique within a document. Also, the Add, Replace, and Delete elements require a "target" element. The sample journal xmiCAS shown in Section 1.3.2 is not a valid XMI document since the Add elements are missing the "target" element and the Replace elements use the same ID.

1.4.2 Queries for Inspecting Journal xmiCAS
After the journal xmiCAS is saved, we will need to address the issue of what queries are useful for inspecting the journal. Which queries are most interesting to implement?

Query examples are:

- examine all the changes made by a specific AE
- examine the change history of a specific feature structure
- get an ordered list of AEs run on the CAS
- see which FS were modified by downstream AEs
- track the creation/modification of specific types

2. UIMA Component Data Collection

2.1 Use-case
When UIMA applications have been deployed at the customer site and one of the components is acting badly, it is important to have a tool that can capture all relevant input data and control parameters which developers can then use to reproduce the problem.

2.2 Proposal
Provide a mechanism to collect all data controlling the behavior of a component deployed within an aggregate, and the ability to later replay these data through the problematic AE in a stand-alone environment.

Details on the proposal includes.

- Data Collection
  - full CAS definition of type system, type priorities, FS indexes
  - result specification
  - configuration parameter settings
  - input CAS(es)

- Reproducing the problem
  - A new application driver replays the collected data through the specified AE.

3. Operational Descriptor XML

3.1 Proposal:
Provide a format for specifying the parameters to control the behavior of the AEs for:

- provenance tracking of UIMA CAS content
- component data collection

3.2 Samples
The following shows some samples of the Operational Descriptor XML.

3.2.1 Sample for Provenance Tracking of All AEs:

```
    <operationalDescription>
        <journal>
            <description>
               An example of journalizing "all" delegate AEs and "all" types.
            </description>

            <topDescriptor>
               <import location="deploy/as/MeetingFinderAggregate.xml"/>
            </topDescriptor>

            <!-- list of analysis engines to journalize -->
            <targetAnalysisEngines all=true />  <!-- select all AEs-->

            <!-- list of types to focus on -->
            <targetTypes type="uima.cas.TOP"/>  <!-- select all types -->

            <!-- where to save journal files -->
            <output directory="outputDirectory" />
        </journal>
    </operationalDescription>
```

3.2.2 Sample for Provenance Tracking of a Single AE:

```
    <operationalDescription>
        <journal>
            <description>
                An example of journalizing the "DateTime" delegate AE in the "MeetingDetectorTAE" aggregate
                and focusing on "DateAnnot" and "TimeAnnot" types. The journal files will be saved in the
"outputDirectory" directory.
            </description>

            <topDescriptor>
               <import location="deploy/as/MeetingFinderAggregate.xml"/>
            </topDescriptor>

            <!-- list of analysis engines to journalize -->
            <targetAnalysisEngines>
                <analysisEngine key="MeetingFinderAggregate/MeetingDetectorTae/DateTime" />
            </targetAnalysisEngines>

            <!-- list of types to focus on -->
            <targetTypes>
                <type>org.apache.uima.tutorial.DateAnnot</type>
                <type>org.apache.uima.tutorial.TimeAnnot</type>
            </targetTypes>

            <!-- where to save journal files -->
            <output directory="outputDirectory" />
        </journal>
    </operationalDescription>
```

3.2.3 Sample for Data Collection:

```
<operationalDescription>
    <dataCollectionParameters>
        <description>
          The following is an example of collectiong data related to the problematic "Meeting" delegate AE
          in the "MeetingDetectorTAE" aggregate. The data files will be saved in the "outputDirectory"
directory.
        </description>
        <topDescriptor>
            <import location="deploy/as/MeetingFinderAggregate.xml"/>
        </topDescriptor>

        <!-- list of analysis engines targeted for data collection -->
        <targetAnalysisEngines>
            <analysisEngine key="MeetingFinderAggregate/MeetingDetectorTae/Meeting" />
        </targetAnalysisEngines>

        <!-- where to save collected data -->
        <output directory="outputDirectory" />
    </dataCollectionParameters>
</operationalDescription>
```