DARPA Communicator XML Log Standard

John Aberdeen

The MITRE Corporation

aberdeen@mitre.org

Presented to the W3C

Voice Browsers Working Group

26 September 2000

MITRE

History and Overview (1)

- In support of evaluation, MITRE has developed an XML log file standard
- In October of 1999 the Communicator Evaluation Subcommittee met to begin developing a set of metrics that we wanted to collect to evaluate our systems
- We've been calling these metrics DMAs (Definition, Motivation, Algorithm)
- MITRE developed a specification for annotating the XML logs with attributes necessary for calculating the DMA metrics, and a suite of tools for manipulating and scoring the logs
  - XML rule-based annotation framework
  - Log review, manipulation and scoring tool (Python)

History and Overview (2)

- Communicator sites participating in the evaluation (AT&T, BBN, University of Colorado, CMU, IBM, Lucent, MIT, MITRE, and SRI) began collecting logs in log standard format, and submitted sample logs to MITRE
- MITRE worked with the participating sites to ensure that their logs were compliant with the standard
- The entire process has driven improvements in the log standard and the log manipulation and scoring tools

Log File Format

- Example log fragment

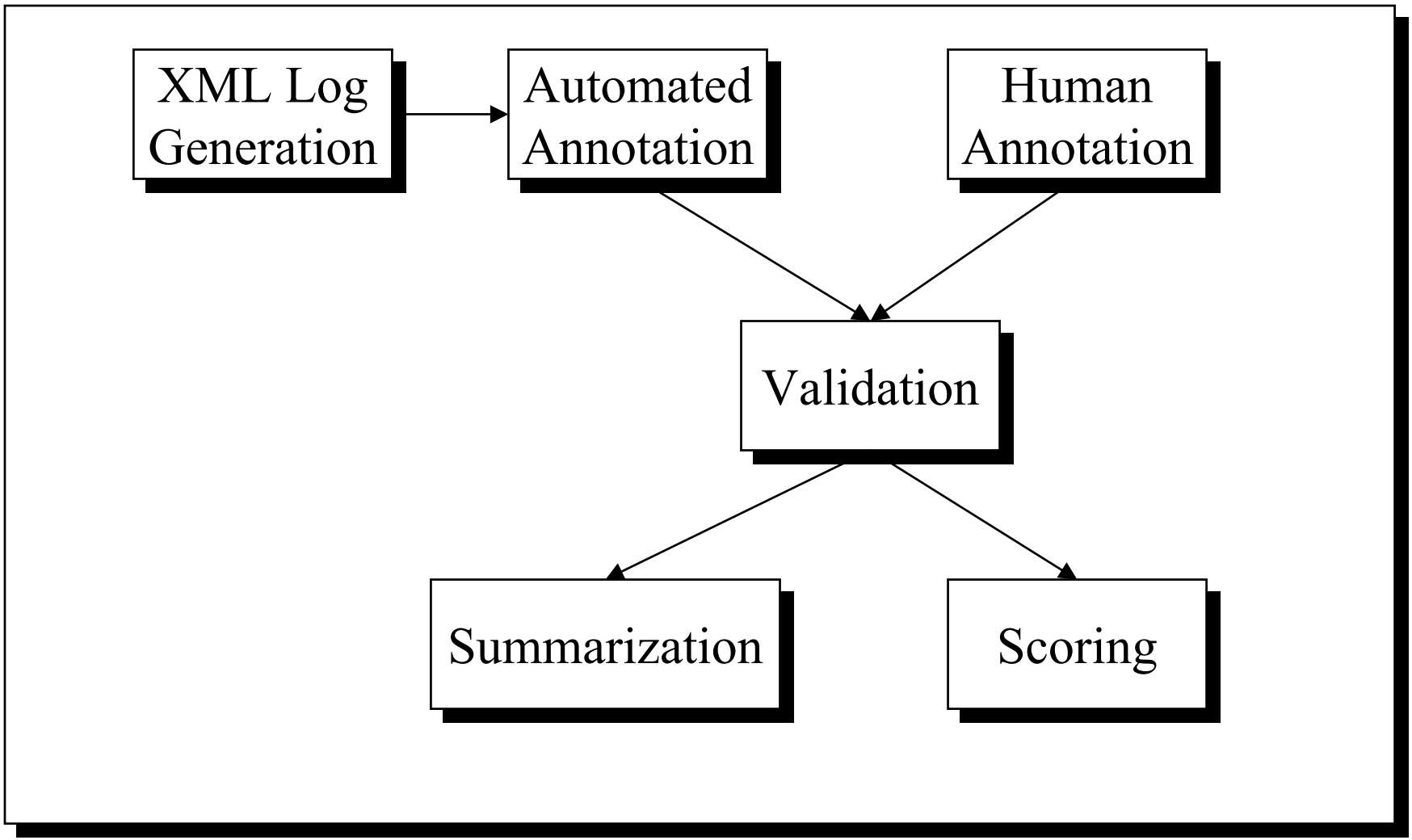
```
<GC_LOG logfile_version="travel, version 2.0 cfone">
 <GC_SESSION etime="941473494.820000" stime="941473394.650000" id="199.94.106.6:20300:0">
  <GC_TURN etime="941473406.620000" stime="941473394.650000" id="-1">
   <GC_OPERATION server="nl" type_new_turn="system" tidx="3" name="paraphrase_reply"
   reply_status="normal" etime="941473394.690000" stime="941473394.660000"
   location="localhost:11000" turnid="-1">
    <GC_DATA key=":reply_string" dtype="string" direction="out">
     Hi! Welcome to Mitre's Travel demonstration. This call is being recorded for system development.
     You may hang up or ask for help at any time. How can I help you?
    </GC_DATA>
   </GC_OPERATION>
   ...
```

- XML
  - There are a growing number of viewers available, as well as a variety of parsers in many programming languages

A Valid Log is Two Separate Files

- Automatically annotated XML log
  - There are three ways to generate an annotated XML log
    - Use MITRE's tools to translate a raw MIT log to a raw XML log, then use MITRE's tools to apply rules that automatically annotate the log (many sites did this)
    - Generate the raw XML log directly, and use MITRE's tools to apply rules that automatically annotate
    - Generate the annotated log directly (one site did this)
  - Scoring and validation tools do not care which method is used to generate the annotated log
- Human annotations XML file
  - Human annotations (transcriptions, task completion) are done offline and kept in a separate file to facilitate alternate annotations (can be later integrated)

The Log Evaluation Process



XML Automated Annotation

- Uses a declarative rules file to automatically add annotations necessary for DMA metrics calculations
- The rules look for "landmarks" in the raw XML log, and add the necessary attributes
- After the rules file is written (once for each system) XML annotation is a fully automatic process

```
<RULE>
   <GC_OPERATION name="nop"
                          new:type_new_turn="user">
      <GC_DATA key=":listening_has_begun"/>
   </GC_OPERATION>
</RULE>
```

Sample Summarization Output

Tue Jun 8 1999 at 16:23:05.26 to Tue Jun 8 1999 at 16:23:05.85: Overall task started.

Tue Jun 8 1999 at 16:23:16.99: Task-specific portion started.

Tue Jun 8 1999 at 16:26:46.05: Task ended.

Task completion status: completed.

Tue Jun 8 1999 at 16:23:05.21 to Tue Jun 8 1999 at 16:23:05.25: New system turn began.

Tue Jun 8 1999 at 16:23:06.57: System started speaking.

Tue Jun 8 1999 at 16:23:16.81: System finished speaking.

System said: Hi! Welcome to Mitre's Travel demonstration. This call is being recorded for system development. You may hang up or ask for help at any time. How can I help you?

Tue Jun 8 1999 at 16:23:16.99: New user turn began.

Tue Jun 8 1999 at 16:23:17.55: User started speaking.

Tue Jun 8 1999 at 16:23:25.59: User finished speaking.

Recognizer heard: is please and i+d like to i want the earliest flight from what time from new york to washington tomorrow

User said: {breath yes please I'd like to %uh book a flight from Wa- %uh from New York to Washington tomorrow

...

XML Scoring

- Reads the annotated XML file as well as the human annotations XML file, and produces a score report for the DMA metrics
- Metrics calculated
  - Task completion
  - Durations
    - on-task duration, total task duration, response latency, mean system utterance duration, mean system turn duration
  - Counts
    - turns to task end, mean user words/turn, mean system words/turn, error messages, help messages, user words to task end, system words to task end, number of reprompts

Sample Scorer Output

- Sample data from 1 call
- Actual output is an HTML table, with 1 row for each call

Task completed	On-task duration (secs)	Total task duration (secs)	Turns to task end	Mean user words/turn	Mean system words/turn	Error messages
1	176.81	202.68	21	4.20	19.70	0

Help messages	Response latency (secs)	User words to task end	System words to task end	Number of reprompts	Mean system utterance duration	Mean system turn duration
0	5.63	42	197	0	6.04	9.81

Metrics that We Still Need to Define

- Semantic accuracy and everything that depends on it
  - Mean User Concepts per Turn
  - Mean Concept Efficiency
  - User Repeats
  - State of Itinerary
- During FY01, the participating sites will be developing new metrics that relate to a systems ability to support mixed initiative in complex tasks
  - The Evaluation Subcommittee (or perhaps the Communicator Advisory Committee) will review these new metrics for inclusion in subsequent evaluations
  - The XML log standard will evolve to support these new metrics

Conclusion

- We've had success with the log standard in the Communicator evaluation
- The log standard will evolve as the Communicator evaluations evolve
- The current XML log standard is NOT tailored to any particular domain, but it IS currently tailored to the Communicator Program
- The XML log standard is being actively used by participants in the Communicator Program
- Further details about the log standard and the DMA implementations are available at:

  http://fofoca.mitre.org/logstandard/index.html