Improving the Reproducibility of PAN’s Shared Tasks

Bauhaus-Universität Weimar
Universitat Politècnica de València
Autoritas Consulting
University of the Aegean

Martin Potthast, Tim Gollub, Benno Stein
Paolo Rosso
Francisco Rangel
Efstathios Stamatatos

[pan.webis.de]
Improving the Reproducibility of PAN’s Shared Tasks

Outline
· About Shared Tasks
· The TIRA experiment platform
· Plagiarism Detection, Author Profiling, and Author Identification
· Summary
About Shared Tasks
About Shared Tasks

Terminology

The term “shared task” refers to computer science events that invite researchers and practitioners to work on a specific problem of interest, the task.*

Goals

- development of new theories / approaches
- implementation of suited softwares
- evaluation of currently achievable performance

*Typical terms used in this regard are: campaign, challenge, competition, contest, or cup.
About Shared Tasks

Terminology

The term “shared task” refers to computer science events that invite researchers and practitioners to work on a specific problem of interest, the task.*

Goals

- development of new theories / approaches
- implementation of suited softwares
- evaluation of currently achievable performance

Pros

- task standardization
- evaluation resource development
- transfer from academia to industry

Cons

- “task concentration” (less diversity)
- winner imitation
- repeated participation fatigue

*Typical terms used in this regard are: campaign, challenge, competition, contest, or cup.
About Shared Tasks

Terminology

The term “shared task” refers to computer science events that invite researchers and practitioners to work on a specific problem of interest, the task.*

Goals

- development of new theories / approaches
- implementation of suited softwares
- evaluation of currently achievable performance

Pros

- task standardization
- evaluation resource development
- transfer from academia to industry

Cons

- “task concentration” (less diversity)
- winner imitation
- repeated participation fatigue

Success indicators

- participation (registrations, downloads, submissions)
- scientific impact (citations)

*Typical terms used in this regard are: campaign, challenge, competition, contest, or cup.
About Shared Tasks

Timeline of Shared Tasks in Human Language Technologies

- **Speech recognition**, **Multimedia processing**
- **Natural language processing**, **Machine translation**, **Named entity recognition**
- **Text synthesis**, **Question answering**
- **Information retrieval**, **Text mining**
- **Information extraction**, **Semantic analysis**, **Topic detection and tracking**
- **Natural language processing**, **Machine translation**, **Named entity recognition**
- **Speech recognition**, **Multimedia processing**
### About Shared Tasks

#### Shared Tasks by Submission Type

**Corpus** (and what may be published to participants)

<table>
<thead>
<tr>
<th>Training data</th>
<th>Training data ground truth</th>
<th>Test data</th>
<th>Test data ground truth</th>
</tr>
</thead>
</table>

**Software** (and what may be submitted by participants)

<table>
<thead>
<tr>
<th>Software source</th>
<th>Software executable</th>
<th>Software run</th>
</tr>
</thead>
</table>
# About Shared Tasks

## Shared Tasks by Submission Type

<table>
<thead>
<tr>
<th>Corpus (and what may be published to participants)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
<td>Training data ground truth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software (and what may be submitted by participants)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Software source</td>
<td>Software executable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Submission type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant</td>
<td>Organizer</td>
</tr>
</tbody>
</table>

Run submission
### About Shared Tasks

#### Shared Tasks by Submission Type

**Corpus** (and what may be published to participants)

<table>
<thead>
<tr>
<th>Training data</th>
<th>Training data ground truth</th>
<th>Test data</th>
<th>Test data ground truth</th>
</tr>
</thead>
</table>

**Software** (and what may be submitted by participants)

<table>
<thead>
<tr>
<th>Software source</th>
<th>Software executable</th>
<th>Software run</th>
</tr>
</thead>
</table>

#### Submission type

<table>
<thead>
<tr>
<th>Participant</th>
<th>Organizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant</td>
<td>Organizer</td>
</tr>
</tbody>
</table>

Run submission

Managed software submission
# About Shared Tasks

## Shared Tasks by Submission Type

### Corpus (and what may be published to participants)

<table>
<thead>
<tr>
<th>Training data</th>
<th>Training data ground truth</th>
<th>Test data</th>
<th>Test data ground truth</th>
</tr>
</thead>
</table>

### Software (and what may be submitted by participants)

<table>
<thead>
<tr>
<th>Software source</th>
<th>Software executable</th>
<th>Software run</th>
</tr>
</thead>
</table>

### Submission type

- **Run submission**
- **Managed software submission**
- **Participant-in-charge software submission**
## About Shared Tasks

### Shared Tasks by Submission Type

<table>
<thead>
<tr>
<th>Corpus (and what may be published to participants)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software (and what may be submitted by participants)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software source</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Submission type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run submission</td>
</tr>
<tr>
<td>Managed software submission</td>
</tr>
<tr>
<td>Participant-in-charge software submission</td>
</tr>
</tbody>
</table>

- **PAN 2009-2011**: run submission
- **PAN 2012**: managed software submission (1 task)
- **PAN 2013**: managed software submission (all tasks)
- **PAN 2014**: participant-in-charge software submission via TIRA
The TIRA experiment platform
The TIRA experiment platform
Software Submission Challenges ➔ Approaches

1. Environment diversity ➔ virtualization
   Support a wide variety of programming languages and operating systems.

2. Executing untrusted software ➔ virtualization
   Better be safe than sorry when executing binaries from a third party.

3. Data leakage ➔ sandboxing
   Prevent data leaking by running software in a secured environment.
The TIRA experiment platform
Software Submission Challenges ➔ Approaches

1. Environment diversity ➔ virtualization
   Support a wide variety of programming languages and operating systems.

2. Executing untrusted software ➔ virtualization
   Better be safe than sorry when executing binaries from a third party.

3. Data leakage ➔ sandboxing
   Prevent data leaking by running software in a secured environment.

4. Error handling ➔ user interface, unit testing
   Give participants the tools to find and fix their software’s errors.

5. Responsibility ➔ user interface
   Put participants back in charge of their submitted software.
The TIRA experiment platform
Software Submission Challenges ➔ Approaches

1. Environment diversity ➔ virtualization
   Support a wide variety of programming languages and operating systems.

2. Executing untrusted software ➔ virtualization
   Better be safe than sorry when executing binaries from a third party.

3. Data leakage ➔ sandboxing
   Prevent data leaking by running software in a secured environment.

4. Error handling ➔ user interface, unit testing
   Give participants the tools to find and fix their software’s errors.

5. Responsibility ➔ user interface
   Put participants back in charge of their submitted software.

6. Execution cost ➔ provide hardware or raise usage fees
   We provide servers to host virtual machines.
The TIRA experiment platform

System Architecture: User Interfaces

Diagram showing the system architecture with TIRA, VM host, restricted access, corpora, runs, and results.
The TIRA experiment platform
System Architecture: User Interfaces

[Diagram showing the system architecture with interfaces and connections]

Participant

Organizer

© pan.webis.de 2014
The TIRA experiment platform

System Architecture: User Interfaces

- Participant
- Organizer
- TIRA
  - Runs
  - Corpora
  - Results
- VM
  - VM host
  - Software
    - Restricted access
  - XOR
- VNC
- SSH
- Task market
- http://www.tira.io
- Software and runs
- Execution progress
- Run
- Corpora Runs Results
- Task market
The TIRA experiment platform
System Architecture: User Interfaces
The TIRA experiment platform

Demo

TIRA — Experiments as a Service
Improving the replicability of shared tasks in computer science

Browse tasks
Learn more

Image source
The TIRA experiment platform

Log Analysis

Runs on:
- Test data
- Training data

Day

# runs on platform

Official platform release
Deadline approaching reminder
Official deadline

©pan.webis.de 2014
The TIRA experiment platform

Log Analysis

New success indicator for shared tasks

- participant engagement (real-time, personalized)
Summary
The PAN Competition

PAN is a network around digital text forensics.

Mission
- Foster research and development in our tasks
- Push the limits of evaluating them
- Improve methodology for lab-style evaluations

Tasks
- Author Profiling (Given a document, what are its author’s demographics?)
- Author Identification (Given a document, who wrote it?)
- Plagiarism Detection (Given a document, is it an original?)
## Summary

The PAN Competition

<table>
<thead>
<tr>
<th>Statistics</th>
<th>ALLC</th>
<th>SEPLN</th>
<th>FIRE</th>
<th>CLEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task(s)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Follower</td>
<td>78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Registrations</td>
<td>11</td>
<td>21</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Runs/Software</td>
<td>13</td>
<td>14</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Notebooks</td>
<td>8</td>
<td>11</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Attendees</td>
<td>5</td>
<td>18</td>
<td>6</td>
<td>30</td>
</tr>
</tbody>
</table>

**Take-away messages**

- Shared tasks are understudied
- Most shared tasks invite run submissions
- Software submissions feasible at scale iff assisted by technology
- TIRA is the first platform to handle software submissions at scale
Summary
The PAN Competition

### Statistics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task(s)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Follower</td>
<td>78</td>
<td>151</td>
<td>181</td>
<td>232</td>
<td>286</td>
<td>302</td>
<td>110</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Registrations</td>
<td>11</td>
<td>21</td>
<td>6</td>
<td>12</td>
<td>16</td>
<td>53</td>
<td>52</td>
<td>68</td>
<td>110</td>
<td>103</td>
</tr>
<tr>
<td>Runs/Software</td>
<td>13</td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>27</td>
<td>27</td>
<td>48</td>
<td>58</td>
<td>57</td>
</tr>
<tr>
<td>Notebooks</td>
<td>8</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>22</td>
<td>22</td>
<td>34</td>
<td>47</td>
<td>36</td>
</tr>
<tr>
<td>Attendees</td>
<td>5</td>
<td>18</td>
<td>6</td>
<td>30</td>
<td>50</td>
<td>25</td>
<td>36</td>
<td>61</td>
<td>58</td>
<td></td>
</tr>
</tbody>
</table>

### Take-away messages
- Shared tasks are understudied
- Most shared tasks invite run submissions
- Software submissions feasible at scale iff assisted by technology
- TIRA is the first platform to handle software submissions at scale

Thank you for your attention!
TIRA’s User Interfaces
Software and Runs

Virtual Machine
- Operating System: Ubuntu (64 bit)
- RAM: 4GB
- CPUs: 1
- State: running (since 2014-06-22 09:00:00)
- Sandbox state: publicly accessible
- Host: example.com
- SSH Port: 44400
- RDP Port: 55501

Add software | Shutdown | Power off

Software 1
- Command: ./mySoftware -i $inputData -o $outputDir
- Input data: training-data
- Input run: none
- Working directory: /home

Save | Delete | Run

Evaluation
- Measures: precision, recall, accuracy
- Input run: software1 2014-06-22-12-00-00 test-corpus

Run

Runs
<table>
<thead>
<tr>
<th>Software</th>
<th>Run</th>
<th>Input data</th>
<th>Input run</th>
<th>Runtime</th>
<th>Size</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>evaluation</td>
<td>2014-06-22-12-10-00</td>
<td>test-data</td>
<td>2014-06-22-12-10-00-00</td>
<td>00:02:04</td>
<td>24K</td>
<td></td>
</tr>
<tr>
<td>software1</td>
<td>2014-06-22-12-00-00</td>
<td>test-data</td>
<td>none</td>
<td>00:01:54</td>
<td>2.2M</td>
<td></td>
</tr>
<tr>
<td>software1</td>
<td>2014-06-22-11-00-00</td>
<td>training-data</td>
<td>none</td>
<td>00:01:54</td>
<td>2.2M</td>
<td></td>
</tr>
<tr>
<td>software1</td>
<td>2014-06-22-10-00-00</td>
<td>training-data</td>
<td>none</td>
<td>00:00:30</td>
<td>1.1M</td>
<td></td>
</tr>
</tbody>
</table>

Execution Progress

Virtual Machine
- Operating System: Ubuntu (64 bit)
- RAM: 4GB
- CPUs: 1
- State: running (since 2014-06-22 09:00:00)
- Sandbox state: sandboxed
- Host: example.com
- SSH Port: 44400
- RDP Port: 55501

Add software | Shutdown | Power off

Software Running
You started a software on your virtual machine. Only one software can be started at a time. Therefore, access to this control panel is limited until the software is finished. Depending on its type, the size of the input data involved, and the software’s performance characteristics, the completion of this process may take some time.

Software: software1
- Command: ./mySoftware -i $inputData -o $outputDir
- Input data: test-data
- Input run: none
- Run: 2014-06-22-12-00-00
- State: running
- Runtime: 0:00:30
- Last output: 2014-06-22 12:00:30
- RAM used: 3127 MB
- CPU load: 98.00%
TIRA’s User Interfaces

Run Evaluation Run (excerpt)

```
python shared-task-evaluation.py -i alice/2014-06-22-12:00-00/output -t test-data -o /tmp/2014-06-22-12-10-00/output/evaluation.txt
```

"precision": "XXX"
"recall": "XXX"

Note: The output of evaluation runs on test corpora is blinded by default. A task moderator will decide whether to make the results visible.

Siderr

File list

```
test-data/alice/2014-06-22-12:00-00/output
   | [ 257]
   | output1.xml
   | output2.xml

[ 90] output517.xml
   | [ 255]
   | output518.xml
```

0 directories, 518 files
TIRA’s User Interfaces

Task Review

Run Review

Particpant Review

Run Details

Overview
Software: evaluation
Run: 2014-06-22-12-10-00
Input data: test-data
Input run: 2014-06-22-12-00-00
Downloadable: false
Runtime: 00:06:04 (h:mm:ss)
Runtime details:
- System: 7.04 (kernel: 4.12.10, osrelease: 0.4)
- CPU: 2 cores, 512 MB
- Memory: 1 GB
- Disk: 16 GB
- Network: 100 Mbit/s
Size: 24K (15442 bytes)
Lines: 36
Files: 2
Directories: 0

Review
This run has not been reviewed, yet.

Reviewer: Bob
Errors:
- No errors
- Missing output
- Extra output
- Invalid output
- Missing messages in stdout or stderr
- Other kinds of errors: please describe them in the comment below.

Comment

Submit

Stdout
python shared-task-evaluation.py -i alice/2014-06-22-12-00-00/output -t test-data -o /tap/2014-06-22-12-10-00/output/evaluation.txt

"precision": "0.90031"
"recall": "0.57293"

Stderr

File list
test-data/alice/2014-06-22-12-10-00/output/
  [ 246] evaluation.prototext
  [ 108] evaluation.txt
0 directories, 2 files

Download
# TIRA’s User Interfaces

## Evaluation Results Review

### Evaluation Results (published)

<table>
<thead>
<tr>
<th>User</th>
<th>Software</th>
<th>Precision</th>
<th>Recall</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>software1</td>
<td>0.90981</td>
<td>0.67283</td>
<td>00:04:17</td>
</tr>
<tr>
<td>Carol</td>
<td>software3</td>
<td>0.85744</td>
<td>0.29081</td>
<td>00:00:56</td>
</tr>
<tr>
<td>Dan</td>
<td>software2</td>
<td>0.96022</td>
<td>0.84248</td>
<td>00:18:32</td>
</tr>
<tr>
<td>Dan</td>
<td>software3</td>
<td>0.96007</td>
<td>0.84511</td>
<td>00:18:32</td>
</tr>
<tr>
<td>Eve</td>
<td>software1</td>
<td>0.82882</td>
<td>0.84156</td>
<td>00:05:18</td>
</tr>
<tr>
<td>Frank</td>
<td>software10</td>
<td>0.92522</td>
<td>0.81819</td>
<td>00:02:40</td>
</tr>
<tr>
<td>Malory</td>
<td>software1</td>
<td>0.87171</td>
<td>0.91539</td>
<td>00:03:37</td>
</tr>
<tr>
<td>Oscar</td>
<td>software1</td>
<td>0.92707</td>
<td>0.88816</td>
<td>00:00:31</td>
</tr>
<tr>
<td>Peggy</td>
<td>software3</td>
<td>0.90032</td>
<td>0.86267</td>
<td>00:22:10</td>
</tr>
<tr>
<td>Sybil</td>
<td>software2</td>
<td>0.90770</td>
<td>0.79531</td>
<td>00:00:31</td>
</tr>
<tr>
<td>Sybil</td>
<td>software4</td>
<td>0.89179</td>
<td>0.86590</td>
<td>00:00:31</td>
</tr>
<tr>
<td>Trent</td>
<td>software0</td>
<td>0.86606</td>
<td>0.91984</td>
<td>00:00:31</td>
</tr>
</tbody>
</table>

### Evaluations on test-corpus

<table>
<thead>
<tr>
<th>User</th>
<th>Software</th>
<th>Evaluation</th>
<th>Input run</th>
<th>Precision</th>
<th>Recall</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>software1</td>
<td>2014-06-22-12-10-00</td>
<td>2014-06-22-12-00-00</td>
<td>0.90081</td>
<td>0.67283</td>
<td></td>
</tr>
<tr>
<td>Carol</td>
<td>software3</td>
<td>2014-06-15-17-35-38</td>
<td>2014-06-15-17-35-38</td>
<td>0.85744</td>
<td>0.29081</td>
<td></td>
</tr>
<tr>
<td>Dan</td>
<td>software2</td>
<td>2014-06-16-17-17-21</td>
<td>2014-06-16-17-17-21</td>
<td>0.96022</td>
<td>0.84248</td>
<td></td>
</tr>
<tr>
<td>Dan</td>
<td>software3</td>
<td>2014-06-23-20-43-39</td>
<td>2014-06-23-20-17-48</td>
<td>0.96007</td>
<td>0.84511</td>
<td></td>
</tr>
<tr>
<td>Dan</td>
<td>software1</td>
<td>2014-06-16-18-03-43</td>
<td>2014-06-16-17-21-44</td>
<td>0.96243</td>
<td>0.83473</td>
<td></td>
</tr>
<tr>
<td>Eve</td>
<td>software1</td>
<td>2014-06-01-12-52-02</td>
<td>2014-08-21-05-56-23</td>
<td>0.82882</td>
<td>0.84156</td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>software10</td>
<td>2014-06-23-13-31-42</td>
<td>2014-06-23-13-24-21</td>
<td>0.92522</td>
<td>0.81819</td>
<td></td>
</tr>
<tr>
<td>Malory</td>
<td>software1</td>
<td>2014-06-20-23-28-21</td>
<td>2014-06-17-09-28-40</td>
<td>0.87171</td>
<td>0.91539</td>
<td></td>
</tr>
<tr>
<td>Oscar</td>
<td>software1</td>
<td>2014-06-19-00-54-42</td>
<td>2014-08-18-23-50-04</td>
<td>0.92707</td>
<td>0.88816</td>
<td></td>
</tr>
<tr>
<td>Peggy</td>
<td>software3</td>
<td>2014-06-23-05-35-34</td>
<td>2014-08-22-03-33-32</td>
<td>0.90032</td>
<td>0.86267</td>
<td></td>
</tr>
<tr>
<td>Sybil</td>
<td>software2</td>
<td>2014-06-23-02-56-09</td>
<td>2014-06-22-02-49-41</td>
<td>0.90770</td>
<td>0.79531</td>
<td></td>
</tr>
<tr>
<td>Sybil</td>
<td>software4</td>
<td>2014-06-22-16-55-56</td>
<td>2014-06-22-16-49-05</td>
<td>0.89179</td>
<td>0.86590</td>
<td></td>
</tr>
</tbody>
</table>