translate5

translate5 - Towards a General-Purpose MT Evaluation Tool

Marc Mittag, MittagQI – Quality Informatics, Mössingen
translate5: Open source coming from translation industry

- At first: VM-distribution
- Origin and development history; about my background
- Current use cases in translation industry
  - Proofreading with
    - Terminology recognition
    - With MQM-based error annotation and Quality categorization
    - Setting of Status
    - Simple workflow with mailing of changed segments
  - Terminology correction in large Translation Memories
    - Detailed analysis of correction needs
Created traction in MT-research

- Column-based approach on data
- SQL back-end
- Ability to serve as a hub
  - Ideal to dock all the various tools of MT-research to one hub
    - to make them easily usable for everyone
    - ensure, they won't get lost, if their developer moves on
    - optionally Integrate them in translate5 installer and updater
  - For working on column-based data
  - Helps to bring the MT-research tools to application in the industry
- Continuously developed Open Source Quality assurance tool backed by the industry
- Synergy through usage of common Open Source tool
  - Big potential in bringing MT-research-ideas to translation industry
  - And translation industry perspectives to MT-research
translate5 installation and update

• Possible on Linux and Windows with a few clicks
  http://confluence.translate5.net/display/TIU/Installation+and+update

• On Mac similar to Linux for experienced users

• On local student systems: Installation via preconfigured Ubuntu Virtualbox instance
  http://www.translate5.net/downloads/translate5-ubuntu-virtualbox-preinstalled.zip (coming soon)

Update via
  sudo -u www-data ./install-and-update.sh
Current usage in MT-research

- Live example at http://translate5-metashare.dfki.de/
- Live example: Annotating multiple target alternatives in translate5
  - Download test data via http://confluence.translate5.net/display/BUS
    - CSV example with custom MQM issue types
    - CSV example with multiple target alternatives
Example of serving as a hub for MT-research
Integration of MT-ComparEval

- Write small translate5-plugin
- Optionally integrate MT-ComparEval in installation and update routine
- Show live, where integrate would make sense
Current main features from user perspective

Showing in live example with people trying themselves

- big amounts of data (files and segments)
- Flexible filtering and sorting of columns
- Flexible columns (show specific columns to specific users)
- Multiple import formats
- Commenting on segments
- MQM-tagging
- TermTagging
- User management
- Workflow
- Task overview
- Locking of specific segments
Future application in MT-research

• Serving as a constant database-driven hub
  – to evaluate MT-data with different existing and future research tools, like
    • MT-ComparEval (show it online)
    • BLEU-calculation in a translate5-column
    • Appraise
    • QuEst
    • Word-Aligners
    • Part of Speech Tagger
  – to research and compare existing annotation projects by
    • Copying the data from existing in new task
    • Re-translate data in a new column of an existing task
    • Compare the new column with existing ones, etc.
  – to do semi-automatic MQM-markup
Future application in MT-research and translation industry

• Show mock-ups about
  – TM/MT-integration (currently at the end of this presentation)
  – Word-Aligner (to be added)
  – Semi-automated MQM-annotation with diff (to be added)

• Further development ideas
  – Integrate Keystroke-logging
  – Export / Visualize postediting time
Important architecture concepts

- Event-triggered plug-in system
  (show plugin-examples in source-code and translate5 confluence)

- REST-API
  (show REST-API in source-code and translate5 confluence)

- Column-based segment-database (flexible base numbers of columns by task)
  (show database in phpMyAdmin and Entity-concept in source code)

- Queuing, Management and Load-Balancing of external and internal processes
  including dependency management, etc.
  (show termTagger and segment counting in action)

- Pragmatic object-orientated structure

- Based on
  - Zend Framework, PHP (backend)
  - ExtJs, Javascript (frontend; show ExtJs documentation)
  - MySQL database, other SQL-DB systems possible
Discussion

• Which tools would be interesting to integrate?
• How would the development for them work?
• Which of these proposals make sense for the development project at MT-Marathon?
Visualize word alignment in translate5

Norway's rakfisk: Is this the world's smelliest fish?

Norwegen u.a. rakfisch: Das ist der böse riechendste Fisch der Welt?
### Semi-automatic MQM annotations (Mock-up 1; before MQM-selection)

- On segment save the changed string gets selected automatically
- The MQM-issue popup opens up
- The user can select an issue
  (via mouse, keyboard or keyboard shortcut)
If more than one section changed, after the MQM-assignment for the first section, the next gets selected.

After saving the segment, a diff algorithm automatically selects the changed section in the corresponding non-editable version of the column and tags that, too.
Semi-automatic MQM annotations (Mock-up 3; after segment save)

- On segment save the changed string gets selected automatically
- The MQM-issue popup opens up
- The user can select an issue
  (via mouse, keyboard or keyboard shortcut)

If more than one section changed, after the MQM-assignment for the first section, the next gets selected.

After saving the segment, a diff algorithm automatically selects the changed section in the corresponding non-editable version of the column and tags that, too.
<table>
<thead>
<tr>
<th>TM Name</th>
<th>Language 1</th>
<th>Language 2</th>
<th>Color</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>MyMemory</td>
<td>Deutsch (de)</td>
<td>English (en)</td>
<td>Green</td>
<td>MyMemory</td>
</tr>
<tr>
<td>TM of Client XY</td>
<td>Deutsch (de)</td>
<td>English (en)</td>
<td>Cyan</td>
<td>OpenTM2</td>
</tr>
<tr>
<td>Collected Phrases</td>
<td>English (en)</td>
<td>Deutsch (de)</td>
<td>Orange</td>
<td>MySlow TM</td>
</tr>
<tr>
<td>TM of Client XY</td>
<td>Deutsch (de)</td>
<td>Italianisch (it)</td>
<td>Blue</td>
<td>OpenTM2</td>
</tr>
<tr>
<td>Moses MT</td>
<td>English (en)</td>
<td>Deutsch (de)</td>
<td>Red</td>
<td>Moses MT</td>
</tr>
</tbody>
</table>
translate5 MT/TM-integration

open source initiative

translate5

files

work files

Source text | Pivot language | Target text | Comments
--- | --- | --- | ---
1. This file is a based on a part of the php-online-Dokumentation. It's translation is done by a pretranslation based on a very fast winalign-Project and is not at all of the translation art. It's only purpose is the generation of demo-data for translate5. | | | 
4. Installation and Configuration | Installation und Konfiguration | 
5. Installation on Unix systems | Installation auf Unix-Systemen | 
6. Apache 1.3.x on Unix systems | Apache 1.3.x auf Unix-Systemen | 
7. Apache 2.x on Unix systems
translate5 MT/TM-integration

This file is a part of the php-online-Dokumentation. It's translation is done by a pretranslation based on a very fast winalign-Project and is not at all state of the translation art. Its only purpose is the generation of demo-data for translate5.

1. Apache 2.x on Unix systems - Manual
   Apache 2.0 auf Unixsystemen - Manual

2. PHP Manual
   PHP Manual

3. Installation on Unix systems
   Installation auf Unix-Systemen

4. Installation on Linux systems
   Installation auf Linux-Systemen

5. Installation on a Unix system
   Installation auf einem Unix-System

Installation on PC systems
   Installation auf PC-Systemen