An Open Service Framework For Next Generation Localisation

Stephen Curran
Trinity College, Dublin
Irish government funded research project into localisation

TCD, DCU, UCD, UL

Microsoft, Symantec, IBM, SDL, VistaTec
Research groups

• Language technologies
  • Machine translation, text analytics, speech synthesis and recognition

• Localisation processes and standards

• Systems framework
  • Software integration
  • Service Oriented Architecture for localisation
Presentation agenda

• What is localisation?
• Current localisation process and supporting software
• Problems with this software
• An open service-oriented framework for localisation
• Future work
What is localisation

• Adaption of content to a locale or language

• Product localisation
  • Adaption of a product to sell into another language market

• Not just translation
  • Engineering work
  • Adaption of layout
  • Testing
Industrial partners

• Software localisation
• Content types
  • User interface
  • Product help
  • Printed documentation
  • Marketing content
  • Online user support pages
Localisation process
Translation memory

- Database of previous translations
- No need to translate the same sentence twice
- Exact match/fuzzy match
- Translation cost calculated based on translation memory hits
Translation memory leverage

<table>
<thead>
<tr>
<th>Match types</th>
<th>Words</th>
<th>Percentage payment of full word rate</th>
<th>Equivalent wordcount to be paid at full word rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>× Translated</td>
<td>0</td>
<td>10%</td>
<td>0</td>
</tr>
<tr>
<td>Repetitions</td>
<td>2323</td>
<td>10%</td>
<td>232</td>
</tr>
<tr>
<td>100%</td>
<td>8831</td>
<td>10%</td>
<td>883</td>
</tr>
<tr>
<td>95%-99%</td>
<td>4821</td>
<td>30%</td>
<td>1446</td>
</tr>
<tr>
<td>85%-94%</td>
<td>5578</td>
<td>60%</td>
<td>3347</td>
</tr>
<tr>
<td>75%-84%</td>
<td>11132</td>
<td>60%</td>
<td>6679</td>
</tr>
<tr>
<td>50%-74%</td>
<td>24304</td>
<td>100%</td>
<td>24304</td>
</tr>
<tr>
<td>No Match</td>
<td>35144</td>
<td>100%</td>
<td>35144</td>
</tr>
</tbody>
</table>

Total wordcount: 92133

Total CATCount: 72035
Termbase

- Translation of core terminology or phrases
- Ensures consistent translation of terminology across the product
- Definition of term to help translator
Computer assisted translation tool
Translation management system

• Manages end-to-end process
• Management of:
  • Workflows
  • Users
  • Projects
  • Translation memories and glossaries
Localisation standards

• LISA standards
  • Translation Memory eXchange (TMX)
  • Termbase eXchange (TBX)
  • Segmentation Rules eXchange (SRX)

• OASIS standards
  • Localisation Interchange File Format (XLIFF)
Problems with current software

• Lack of support for data standards
  • Proprietary formats
  • Import but not export

• Lack of standard protocols/interfaces for common tasks

• Inflexible workflows
  • Support for a limited set of activities
New service-oriented approach

• Support for localisation data standards
• Definition of interfaces for common localisation tasks
• Use of open web based protocols and technologies
  • HTTP
  • WSDL and SOAP
  • REST
Initial set of services

• Machine translation
  – Service wrapper for the Moses decoder

• Language classification service
  – Service wrapper for the TextCat library

• Domain classification service
  – Service wrapper for text classification algorithm from academic partner
BPEL

- Compose services into an executable workflow using BPEL
- XML based language
- Drag and drop designers available
- Netbeans and Glassfish
Plug into existing workflows
Service oriented architecture
Future work

• Propose interfaces for common localisation components
• Implement these interfaces wrapping software made available to the project
• Expand on example processes