Real-World Application of a Machine Translation Workflow

Tomáš Fulajtár
MORAVIA IT
Prague, September 7th, 2015
Personal Introduction

• 10+ years in Moravia
  – 8 years of engineering on SW localization
    (all OS platforms, QA work)
  – 2+ years in Language Technology Group
    (managing Moses environment)
  – Focus on MT system training,
    optimizations, technology improvements

Tomáš Fulajtár
tomasfu@moravia.com
+420 545 552 340
Overview

• The Machine Translation Ecosystem
• Localization 101
  – Translation Memory
  – Computer Aided Translation Tools
  – Combine TM + MT
• Challenges in applied MT
  – Moravia experience (solutions)
• Case Studies
• Open Topics
• Summary
The MT Ecosystem

- Academic
- Commercial Development
- In-house MT owners
- LSP (Language Service Providers)
Language Service Providers

- Aka Translation Agencies
- Well established, yet dynamic industry

- Provide numerous linguistic services for clients
- Deal with
  - Variety of customers/products/content (IT, medical, automotive, etc.)
  - Huge number of languages (150+)
  - Wide range of CAT tools
Moravia at a Glance

Global Offices: 10+
Internal Headcount: 800+
Specialist Translators: 5000+

Languages: 120+
CAT tools: 12+
MT Engines: 100+
Growth of MT in production environment

- **MT**: Orange
- **Prediction**: Light blue

Yearly comparison:
- **2013**: Low MT
- **2014**: High MT
- **2015**: High MT with prediction
Localization 101
Localization workflow 101

• Based on Translation Memory (TM) technology developed in the early 1990s (Trados Translator's workbench 1994)
• Database containing segments of translated content
• CAT tools are based on TM technology & concepts
  – Analysis (Weighted Word Count calculated from fuzzy match algorithm)
  – Segmentation
  – Translation
  – Term extraction
  – Concordance
Advantages:

- Improves translation speed (long term projects)
- Savings for large projects
- Ensures better consistency/terminology
- Online collaboration (TM server/cloud)

Obstacles:

- Manual maintenance necessary to keep good quality
- Lower leverage for non-technical domains
- Could not generate long sentences well (the parts not always fit together)
Machine Translation

Advantages:
- Translate unseen content
- Consistent output
- Online translation

Obstacles:
- Quality differs among languages
- Need large corpus
- Non-transferable across domains
Combining Technologies: TM + MT

Goal to match source content against TM to plug in any previously-translated content, applying the MT/PE to only new strings.
Why MT in the LSP industry at all?

Customers interested in:
• On-demand translations for daily content
• High volumes of content but limited budget to human translate everything
• Need for instant translation (e.g., chat)
• Right quality for right content type
• Short turnaround time (TAT) for translation
Challenges of Applied MT
Challenges of Applied MT

• Integrating MT into the localization workflow
  – Connecting to CAT tools for efficient post-editor interaction with MT output

• Measuring the impact of MT on the localization process
  – Measuring Engine quality
  – Measuring post-editor productivity

• The impact of MT on the business relationship
Integrating MT into the localization workflow

mtoservices

• Internal General Purpose Tool
  • Operates on interchangeable loc. Formats (TMX, XLIFF)
  • MT technology Agnostic

• Connectors to CAT tools:
  – Native support (MS Hub, Google etc.)
  – MosesXMT - Moravia LetsMT compatible API (MemSource)
  – LTGear - global MT API (in development)
Importance of TMX

• File format for TM data interchange
• Bilingual
• Used in several scenarios throughout LSP process
  – CAT Tools
    • Integration with MT (for larger volumes)
  – Corpus creation
  – MT Training
Measuring the impact of MT (I)

Moravia Platform for EVAluation of MAchine TranSLation

• Covers both Machine Contribution and Human Contribution to the final translation
  – Machine Contribution : Meteor
  – Human Contribution : RedBall (customized TER)

• Segment Level Testing A/B (SLAB)
Measuring the impact of MT (la) - Score Categorization

Meteor Score Categorization

RedBall TER Score Categorization

Meteor Score Categorization Explained

RedBall TER Score Categorization Explained
Measuring the impact of MT (lb) - Categorization By Segment Length
Measuring the impact of MT (II)

Human Evaluation

• Custom solution - Excel sheet processing
  + Works offline
  – Manual distribution and collection

• Towards standardization:
  – TAUS’s Dynamic Quality Framework (DQF)
  – WMT tools (Appraise)
Sample MT Engagement Process

Preparation
- Corpus Creation
- Training/Tuning
- Evaluation
- Human review

Production
- Extracting MT content
- MT Decoding
- Post Editing

Post-production Analysis
- MT Contribution
- Human Contribution
- Estimated PE Effort
MT Business negotiations

• Challenge - everybody thinks MT is free (e.g., Google)
• We approach MT as a productivity tool
• Important part of delivering a full suite of language services, not a standalone service or simple packaged solution
• Cost is not most important driver - it is speed / shorter turnaround time
  – Pricing is part of the business relationship. MT usage is only one of many driving factors
Case Studies
# Case Study:
## MT Program for Top Backup & Recovery Software Company

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Approach</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Continually cut down on localization costs</td>
<td>• Work with client to understand quality level mapped to content type</td>
<td>• All target content determined to be viable</td>
</tr>
<tr>
<td>• Improve turnaround time</td>
<td>• Develop workflows: MT+PE for tech docs, Raw MT for Knowledge Base</td>
<td>• Initial cost reduction 10%</td>
</tr>
<tr>
<td>• Translate more content</td>
<td>• Select best of breed from multiple MT technologies</td>
<td>• Will increase to 20-30% reduction over time due to incremental retraining</td>
</tr>
<tr>
<td></td>
<td>• MS Translator Hub wins</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Evaluated outputs with automatic (EVAMAT) and human methods</td>
<td></td>
</tr>
</tbody>
</table>
## Case Study: Machine Translation Program for Automotive Diagnostic Tools

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Approach</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Stretch an existing localization budget</td>
<td>• Understand quality level mapped to content type</td>
<td>• SW content didn’t reach the required quality due to terminology issues</td>
</tr>
<tr>
<td>• Improve turnaround time</td>
<td>• Tested Moses and MS Translator Hub; Moses prevailed</td>
<td>• Help + Doc content was viable with PE</td>
</tr>
<tr>
<td>• Translate more content</td>
<td>• Evaluated outputs with automatic (EVAMAT) and human methods</td>
<td>• Initial time reduction minimal</td>
</tr>
<tr>
<td>• Low quality source content</td>
<td>• Created terminology strategy: Moravia automated terminology tool to improve consistency</td>
<td>• Source improvement improved MT output</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Corpus enhancements planned in future</td>
</tr>
</tbody>
</table>
Case Study: Machine Translation and Post-Editing for Software Giant

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Approach</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Reduce global support costs and improve user experience</td>
<td>• MS Translator Hub engine</td>
<td>• MT viable for 29 of 34 target languages, with average TM leverage of 72%</td>
</tr>
<tr>
<td>• Increase target languages, content types and volume</td>
<td>• Optimize TMs to achieve higher leverage</td>
<td>• Total program savings of 10%</td>
</tr>
<tr>
<td>• Map content types to required levels of quality</td>
<td>• Full PE on high-profile content to achieve human quality</td>
<td>• 1.5M machine-translated words produced in 2014</td>
</tr>
<tr>
<td>• Get more value from the existing budget by improving productivity</td>
<td>• Light PE on standard content to achieve acceptable “gisted” quality level</td>
<td>• Increased user self-support in more local markets; global support costs down by 10%+</td>
</tr>
</tbody>
</table>
Open topics

- MT Prediction
- More human-like Automated metrics
- Better MT for under-resourced languages
- Morphologically rich languages challenges
  - Czech, Korean, Finno-Ugric, Turkish
- Adding syntactic features into MT
Summary

• LSP usage of MT poses several challenges
• Presented solutions fitting in LSP environment
• Moravia has developed solutions for:
  – Integration
  – Impact Evaluation and Analysis

• Case studies representing different MT approaches
• Moravia is looking forward to participating in the MT Marathon
Questions
Thank you!

Tomáš Fulajtár

tomasfu@moravia.com

+420 545 552 340