Building A Multi-Purpose MT Portfolio
AGENDA

Multi-Purpose MT?
—
MT usage scenarios and requirements
—
Case Study 1: **Entity Protection**
—
Case Study 2: **Custom Terminology**
—
Case Study 3: **Tone of Voice**
—
Key Takeaways
MULTI-PURPOSE MT?
ENTERPRISES MASSIVELY FAIL TO ADOPT AI

20%* Wrong vendor selected
Failed integrations
Failed pilots
Failed to deliver ROI

* Share of US companies with successful AI deployment (Deloitte State of Cognitive Survey 2017)
© Intento, Inc. / September 2020
BRIDGING THE GAP BETWEEN MT CAPABILITIES AND ADOPTION

MT Procurement

MT Need

Localization

MT Systems
BRIDGING THE GAP BETWEEN
MT CAPABILITIES AND ADOPTION

MT Procurement
__
MT Curation

MT Need

Localization

MT Systems
BRIDGING THE GAP BETWEEN MT CAPABILITIES AND ADOPTION

MT Procurement
___
MT Curation
___
Multi-Engine MT

MT Need

Localization

MT Systems
BRIDGING THE GAP BETWEEN MT CAPABILITIES AND ADOPTION

MT Procurement

MT Curation

Multi-Engine MT

Multi-Purpose MT

MT Need

Localization

Customer Service

Office Productivity

Global Community

MT Systems
Instant ROI on the investments already made

Combining resources of multiple stakeholders to benefit everyone

MT Requirements beyond the objective linguistic quality

Optimizing for features may compromise the quality
MULTI-PURPOSE MT

Localization
- PRODUCT INFORMATION MANAGEMENT
- AGILE LOCALIZATION
- POST-EDITING / TMS
- SUPPORT TICKETS
- LIVE CHATS
- CHAT BOTS

Community
- ON-THE-FLY MT FOR UGC CONTENT
- REAL-TIME USER COMMUNICATION
- COMMUNITY PORTAL TRANSLATION
- ON-THE-FLY MT FOR INTRANET WEBSITES
- OFFICE TOOLS
- MT PORTALS

Customer Service

Office productivity

© Intento, Inc. / September 2020

Proceedings of the 14th Conference of the Association for Machine Translation in the Americas
October 6 - 9, 2020, Volume 2: MT User Track

Page 320
MULTI-PURPOSE MT
REQUIREMENTS BEYOND QUALITY

large text translation
—
batch translation
—
latency and jitter
—
tolerance to bad source
—
language detection
—
tag support

multilingual source
—
profanity control
—
metadata protection
—
entity protection
—
custom terminology
—
tone of voice consistency
ADDITIONAL CHALLENGES WITH SPECIFIC COMBINATIONS

large text translation + HTML support
—
source language detection + multilingual source
—
…
## MT Requirements Matrix

Every case has its own needs.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Post-editing / TMS</th>
<th>Support tickets</th>
<th>Live chats</th>
<th>Chatbots</th>
<th>On-the-fly UGC</th>
<th>Real-time communication</th>
<th>Knowledge bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large text translation</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch translation</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latency and jitter</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tolerance to bad source</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language detection</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag support</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilingual source</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profanity control</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metadata protection</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entity protection</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Custom terminology</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tone of voice control</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### MT Requirements Matrix

**Sample**

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Post-editing / TMS</th>
<th>Support tickets</th>
<th>Live chats</th>
<th>Chatbots</th>
<th>On-the-fly UGC</th>
<th>Real-time communication</th>
<th>Knowledge bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large text translation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch translation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latency and jitter control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bad source language detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag support</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilingual source control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profanity control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metadata protection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entity protection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Custom terminology control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tone of voice control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Also different for inbound and outbound...**
## MT REQUIREMENTS SUPPORT

### BY POPULAR MT ENGINES

<table>
<thead>
<tr>
<th>Feature</th>
<th>Amazon Translate</th>
<th>Google Translate Advanced</th>
<th>DeepL Pro API</th>
<th>IBM Watson Translator</th>
<th>Microsoft Text Translator</th>
<th>ModernMT</th>
<th>Systran PNMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large text translation</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Batch translation</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Latency and jitter</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Bad source language detection</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Tag support</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Multilingual source control</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Profanity control</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Metadata protection</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Entity protection</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Custom terminology</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Tone of voice control</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

- **Supported**
- **Support or its quality depends on the language pair / model**
CASE STUDY 1:
ENTITY PROTECTION
ENTITY PROTECTION
SOME SAMPLES

Simplest cases: protecting email, URLs, phone numbers, file paths
— Crucial for Customer Service
— Easily broken by MT

<table>
<thead>
<tr>
<th>Source text (English)</th>
<th>Machine Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>I just want to let you know about a spam mail I have received on Friday - it's in D:</td>
<td></td>
</tr>
<tr>
<td>\Drv\Prt\Epson\Universal driver x64\ABC6\eeecu120m.inf</td>
<td>Я просто хочу уведомить вас о спаме, который я получил в пятницу - он здесь D:</td>
</tr>
<tr>
<td>\Drv\Prt\Epson\Универсальный драйвер x64\ABC6\eeecu120m.inf</td>
<td></td>
</tr>
<tr>
<td>It has been Ivan Mitrich (ASAP, <a href="mailto:email.some+plus@example.com.tr">email.some+plus@example.com.tr</a>) from Belgrad, but in</td>
<td></td>
</tr>
<tr>
<td>the future it will be me.</td>
<td>Bio je to Ivan Mitrich (ASAP, <a href="mailto:email.some+plus@example.com.tr">email.some+plus@example.com.tr</a>) iz Belgrada, ali u</td>
</tr>
<tr>
<td></td>
<td>budućnosti to ću biti ja.</td>
</tr>
<tr>
<td>Would you like to help with a new phone for the ABC department - (772) 194 59 65 ext</td>
<td></td>
</tr>
<tr>
<td>4406/4408).</td>
<td>Desideri aiutarti con un nuovo telefono per il dipartimento ABC - (772) 194 59 65 ext</td>
</tr>
<tr>
<td></td>
<td>406/4408).</td>
</tr>
<tr>
<td>You must submit such a request via ABC-portal, attached link: <a href="http://www.example.com/en/submit">www.example.com/en/submit</a></td>
<td>Deve inviare o pedido de tal através do ABC-portal, link anexo: <a href="http://www.example.com/pt/submit">www.example.com/pt/submit</a></td>
</tr>
</tbody>
</table>
Selecting the MT based on the default entity protection may compromise the quality.

What if we enforce protection via MT-agnostic NLP?
CASE STUDY 2: CUSTOM TERMINOLOGY
Simplest cases: enforcing acronyms, brand names and other proper nouns.

Without Custom Terminology support, NMT easily breaks them.
CUSTOM TERMINOLOGY IMPROVES FIDELITY

Selecting the MT engine by the custom terminology support may compromise MT Quality

MT-agnostic glossary on a top of NMT
CASE STUDY 3: TONE OF VOICE CONTROL
**TONE OF VOICE CONTROL**

**SAMPLES FROM SUPPORT CHATS**

<table>
<thead>
<tr>
<th>Source text (English)</th>
<th>Machine Translation (German)</th>
<th>COMMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can you share your screen?</td>
<td>Können Sie Ihren Bildschirm freigeben?</td>
<td>FORMAL</td>
</tr>
<tr>
<td>Could you help me?</td>
<td>Kannst du mir helfen?</td>
<td>INFORMAL</td>
</tr>
<tr>
<td>Make sure you report any of</td>
<td>Stellen Sie sicher, dass Sie eines dieser Probleme melden.</td>
<td>FORMAL</td>
</tr>
<tr>
<td>these issues.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can you give an example?</td>
<td>Kannst du ein Beispiel geben?</td>
<td>INFORMAL</td>
</tr>
</tbody>
</table>

Formal vs. Informal

—

Crucial for Live Chats

—

Baseline MT engines are not consistent
TONE OF VOICE CONTROL
DEFAULT MT OUTPUT

English to German
—
210 segments
—
stock models
TONE OF VOICE CONTROL

HOW TO MAKE IT INFORMAL?

Option 1: Use DeepL with formality=less (99.5% accuracy)

Option 2: Generate synthetic training data, hoping translations become more informal

Option 3: MT-agnostic NLP

What if you need a custom model and terminology, or another MT has better linguistic quality for you?

Expensive and time-consuming, also introduces bias into the model

Works to a certain extent, provides a wider choice of MT engines
TONE OF VOICE CONTROL
MT-AGNOSTIC ADJUSTMENT

English to German

210 segments

stock models

let’s make it more INFORMAL
TONE OF VOICE CONTROL
MT-AGNOSTIC ADJUSTMENT

English to German
—
210 segments
—
stock models

let’s make it more
FORMAL
KEY TAKEAWAYS

Multi-Purpose MT brings an instant ROI on the MT investment already made.

Different use cases impose multiple requirements beyond the linguistic quality.

Meeting the requirements takes either the right MT engine choice or clever engineering.

We do both, by implementing MT-agnostic fine tuning algorithms to avoid compromising the MT quality.
THANKS!

ks@inten.to