Company & Technology Overview

Product and Technology Platform Overview

Kirti Vashee – kirti.vashee@asiaonline.net
Asia Online Company Objectives

Revolutionize the Internet experience for non-English speakers in Asia

Revolutionize the enterprise translation process with a comprehensive, continuous learning SMT platform

Provide 1 billion+ local-language pages online using mostly translated open license content, combined with compelling portal and social networking style services in Thailand, Indonesia, India, Malaysia, Philippines, Vietnam and China, Japan & Korea

SaaS environment that allows data cleaning and preparation, develop SMT engines on demand and enable ongoing comprehensive post editing and correction to continuously improve engines

The Consumer Market

The Enterprise Market

Large Buyer & Publisher Perspective

Translation Tools Vendor Perspective

Copyright © 2009, Asia Online Pte Ltd
What makes Asia Online unique?

• The only SMT technology provider that is also a major user of ALT technology on one of the largest translation projects in the world - English Wikipedia (1B Words+) into 11 Asian languages using MT and crowdsourcing.

• The translation tools and technology platform used to accomplish this, is also made available as a SaaS product and onsite servers for the enterprise translation market.

Constant dialogue between the Users and Developers to ensure that rapid translation quality evolution is a reality.
Technology Overview
The Translation Platform Overview

- Data Management
  - Bilingual Data Preparation & Cleaning
  - Bilingual Data Normalization & Optimization
  - Source Cleanup and Preparation
  - Grammar and Spelling validation
  - Monolingual Data Extraction & Analysis

- Hybrid SMT Engine
  - SMT System Training & Development
  - Monolingual Data Training
  - Ongoing Corpus Refinement and Tuning
  - Analysis and Evaluation of N-grams

- Output Proofing and Editing
  - Error Pattern Identification & Correction
  - Automated error correction tools
  - Continuing Cycle of Exception Identification and Correction
  - Development of small sets of new data to correct errors

Copyright © 2009, Asia Online Pte Ltd
Problem: • Only 0.5% of what needs to be translated today is being translated due to cost and time constraints.

Solution: • Machine translation offers, “good enough” quality alternative to human that is faster and lower cost and can deliver huge volumes of content quickly – millions of words per day.
• Where “gold-standard” translations are required, MT + human post editing can deliver the same content at high quality, significantly faster.
Connecting the Language Studio™ Suite

Customer Site
- Project and Resource Management
- Private Data Storage
- Internet
- Asia Online Software as a Service

External Editors and Translators
- Distributed Post Editing
- Translation
- Training
Other language pairs will be created as demand requires or based on customer request.
Pre-Processing Rules
Sentence Segmentation
Word Segmentation
Phrase Reordering
Dates and Numbers
Patterns, Formulas etc.
Pre-Normalization
Spell Checking
Custom Runtime Glossary
Pre-Formatting

Statistical Machine Translation

Post-Processing Rules
Capitalization
Post-Formatting
Grammar Checking
Post-Normalization
XML Tag Reinsertion
Currency Conversion
Cross Referencing
Other custom post processing

Pre-Processing and Post-Processing Rules are implemented in a C# script, which can be customized to fit almost any requirements. This is a true hybrid approach, where rules are used to guide the statistical engine to greater accuracy.
<table>
<thead>
<tr>
<th>Training Data</th>
<th>Preparation</th>
<th>Train</th>
<th>Diagnostic Engine</th>
<th>Train</th>
<th>Version Release</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Foundation Language Pair</td>
<td>13 Cleaning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Foundation Domain</td>
<td>14 Data Generation</td>
<td>20 Error Pattern Analysis &amp; Correction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Custom Data</td>
<td>15 Validate Alignment and Quality</td>
<td></td>
<td>21 Unknown Word Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Bilingual Data</td>
<td>16 Alignment</td>
<td></td>
<td>22 Normalization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Glossaries</td>
<td></td>
<td></td>
<td>23 Error Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Phrase Pairs (TMX, XLIFF, CSV)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Bilingual Documents (MS Word, Plain Text, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Monolingual Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 URLs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Plain Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Tuning Set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Test Set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 Data mining for grammatical patterns Optional (Recommended)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 High Frequency Term &amp; Phrase Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 Normalization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 Post Edit and Train as required</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Optional (Recommended)
The System Interface & Screenshots
Blunt's memoirs revealed little about his spying activities.

Blunt wrote the 30,000-word document after former prime minister Margaret Thatcher exposed his treachery in 1979.

The revelations led to a man who had worked as Surveyor Queen's Pictures being stripped of his knighthood.

His version of events was given to the library in 1984, the year after his death, on condition that it was not displayed for 25 years.

In it, he describes his recruitment by Moscow: "I found that Cambridge had been hit by Marxism and that most of my friends among my junior contemporaries, including Guy Burgess, had joined the Communist Party or were at least very close to it philosophically."

Blunt a écrit le document in 30 000 mot après l'ancien premier ministre Margaret Thatcher avoir exposé son traîtres en 1979.

Les révélations avaient un homme qui avaient fait comme inspecteur des images de la Reine s'arrogant de ses chevauchées.

Son version des événements a été donnée à la bibliothèque en 1984, l'année après son mort, à condition qu'il n'était pas pour 25 ans.

Dans cette fenêtre, il décrit son recrutement de Moscou: "J'ai trouvé que Cambridge a été touché par le marxisme et que la plupart de mes amis au milieu de mes contemporaires de junior - y compris Guy Burgess - soit il a rejoint le parti communiste ou soit il était au moins très proche de ce parti de façon politique."

Blunt's Soviet spying a mistake

Source:
Blunt's memoirs revealed little about his spying activities.

Blunt wrote the 30,000-word document after former prime minister Margaret Thatcher exposed his treachery in 1979.

The revelations led to a man who had worked as Surveyor Queen's Pictures being stripped of his knighthood.

His version of events was given to the library in 1984, the year after his death, on condition that it was not displayed for 25 years.

In it, he describes his recruitment by Moscow: "I found that Cambridge had been hit by Marxism and that most of my friends among my junior contemporaries, including Guy Burgess, had joined the Communist Party or were at least very close to it philosophically."

Target:
Blunt a écrit le document in 30 000 mot après l'ancien premier ministre Margaret Thatcher avoir exposé son traîtres en 1979.

Les révélations avaient un homme qui avaient fait comme inspecteur des images de la Reine s'arrogant de ses chevauchées.

Son version des événements a été donnée à la bibliothèque en 1984, l'année après son mort, à condition qu'il n'était pas pour 25 ans.

Dans cette fenêtre, il décrit son recrutement de Moscou: "J'ai trouvé que Cambridge a été touché par le marxisme et que la plupart de mes amis au milieu de mes contemporaires de junior - y compris Guy Burgess - soit il a rejoint le parti communiste ou soit il était au moins très proche de ce parti de façon politique."

Confidence: 60

Copyright © 2009, Asia Online Pte Ltd
<table>
<thead>
<tr>
<th>Confidence level of the MT output</th>
<th>Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Reviewed</td>
<td>Changed</td>
</tr>
<tr>
<td>Proof Read</td>
<td></td>
</tr>
</tbody>
</table>
Edits are reflected real time in the target document.

Version tracking with visual edits that show the changes made by the user. These changes are used later to build corrective data.
<table>
<thead>
<tr>
<th>No.</th>
<th>Source</th>
<th>Reference</th>
<th>Candidate</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>空気送風機チャンバーは送風機に取付けして動いた。</td>
<td>The chamber for ventilating air provided in a blower through a duct.</td>
<td>An air chamber for the air blower through a duct provided.</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>NPC2は第二キャリアPC2の回転速度を示す。</td>
<td>NPC2 represents the rotational speed of the secondary carrier PC2.</td>
<td>NPC2 second carrier PC2 represents the rotational speed of the device.</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>速歯比を示す。</td>
<td>( M_1 ) is the gear ratio.</td>
<td>( M_1 ) tooth gear ratios are shown.</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>空気用のチャンバーは扇面側と外形形状をしたチャンバーモータと、チャンバー本体の取り付けに合ったノズルで構成。</td>
<td>The chamber for ventilation includes a chamber body having an oval-shaped cross-section, and a nozzle provided in a collective part of the chamber body.</td>
<td>Chamber for ventilation of the oval-shaped cross section and a bottom chamber body 10 screws of a chamber body provided with a nozzle.</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>レンネ方箱は装置の他に組み立てに用いるレンネ方箱構成および構成を模倣する。</td>
<td>The brick dismantling device may also be a brick dismantling device having various other structures.</td>
<td>The brick dismantling device brick dismantling device good even in a variety of other structures.</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>レンネ方箱は装置に各種レンネ方箱が取り付けられた場合でもレンネ方箱の構成、各種レンネ方箱構成を模倣するものである。</td>
<td>The brick dismantling device does not necessarily need to be composed of at least a pair of telescopic rod-shaped members attached to a refractory brick dismantling members at the tip.</td>
<td>A distal end of the brick dismantling device brick dismantling member is attached to at least one pair of telescopic rod-shaped members is that which comprises all necessary.</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>NPC1は第一キャリアPC1の回転速度を示す。</td>
<td>NPC1 represents the rotational speed of the primary carrier PC1.</td>
<td>NPC1 first carrier PC1 represents the rotational speed of the vehicle.</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>ハイパス通過20の角度において切換え20の角度における通過値を設定。</td>
<td>A bypass passage 20 is provided on both sides of the change-over valves 28 30.</td>
<td>A bypass passage 20 of the change-over valves 28, 30 on both sides is provided.</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>前記切換部分における切換え調整機能の変化を示す。</td>
<td>The above notched section can have a cover member for covering the outer side of the above solenoid expansion valve.</td>
<td>The notch can be provided in the cover member for covering the outer side of the solenoid expansion valve.</td>
<td>19</td>
</tr>
</tbody>
</table>
• Training data can often have gaps in coverage and an excess of data in other areas.
• Gaps in coverage reduce translation quality.
• Gaps can quickly be filled via post editing the machine translated output and submitting the data back to the system for further learning.
• Many gaps can be filled with monolingual data only.
• Further gaps can be identified and resolved by analyzing the text that is to be translated for high frequency terms and unknown words
• In some cases incorrect data may be statistically more relevant. Post editing will raise the relevance of the correct grammar.
The best translations come from clean data. Tools determine clean data, but give user the chance to review and remove out of domain data and other unwanted data.
Searches can be constrained and refined by many factors.

Example:

- For Mexican Spanish, set language to Spanish and Country to Mexico.
- For Table (as in Database), set Sites to include “Oracle, Microsoft, ZDNet etc.”
- To avoid overload of one particular style, limit the amount of return data by site.
- Add custom search providers such as existing TMS or public search engines.
The quick brown fox over jumps the lazy dog

The quick brown fox jumps over the lazy dog

Buddha jumps over the wall
Siemens Wind Power CEO jumps over to Repower
Judge jumps over bench in courtroom melee
Military surveillance bot jumps over 25 foot walls
Robbie Maddison jumps over Tower Bridge on motorbike
Man jumps over Grand Canyon
Cow jumps over Moon
With IE9 in sight, Firefox jumps over 50% market share mark
Long jumper Brian Thomas jumps over a car to raise money
Rally car jumps over a crazy fan!
Kobe jumps over a speeding Aston Martin
A deer jumps over a motorcycle
A woman jogging in a California state park jumps over a 100-foot cliff to get away from attacker
An Afghan Army soldier jumps over an irrigation canal while conducting a foot patrol

Language Studio™ Pro analyzes corrections and generates other examples that include the corrected phrase to fill gaps in grammatical patterns. Each post edited correction is amplified producing many other corrective patterns, improving future translations.
Quality Evolution via Corrective Feedback

Key
- Correct
- Mistranslation
- Syntax/Grammar
- Terminology
- Spelling
- Punctuation

Human Feedback
- Correct

Targeted Corrections of Bad Learning
- Correct

Spelling and Terminology
- Correct

Initial System
- Correct

Translation Quality

Human Feedback can raise the raw output to previously unseen quality levels
The highest quality translation engines are specialized on domains.

15 pre-created domains provide a solid foundation upon which high quality customized engines for clients can quickly be created.

A major element and accelerator when implementing stylized translations specific to clients.

This is unique in the industry and gives Asia Online a significant competitive advantage.

Domains can be combined.

<table>
<thead>
<tr>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
</tbody>
</table>
Thank You

kirti.vashee@asiaonline.net

Follow me on Twitter:
@kvashee

Join the Automated Language Translation Group in Linked In to monitor MT discussions

Copyright © 2009, Asia Online Pte Ltd