Better translations with user collaboration -
Integrated MT at Microsoft

Chris Wendt
Microsoft Research
One Microsoft Way
Redmond, WA 98052
christw@microsoft.com

Abstract
This paper outlines the methodologies Microsoft has deployed for seamless integration of human translation into the translation workflow, and describes a variety of methods to gather and collect human translation data. Increased amounts of parallel training data help to enhance the translation quality of the statistical MT system in use at Microsoft. The presentation covers the theory, the technical methodology as well as the experiences Microsoft has with the implementation, and practical use of such a system. Included is a discussion of the factors influencing the translation quality of a statistical MT system, a short description of the feedback collection mechanism in use at Microsoft, and the metrics it observed on its MT deployments.

1 Introduction
Microsoft has deployed its internally developed Statistical Machine Translation system for Microsoft’s own purposes since 2002, mainly for use in the customer support knowledge base and for post-editing software and documentation. Since 2007 the system has been available for use by the general public at the Microsoft/Bing Translator web site. The system includes a mechanism for submitting, rating and approving human quality translations, which are transparently used in subsequent automatic translations as well as MT engine customization and optimization. The experiences with this system are described in the following.

1.1 Microsoft’s statistical MT engine
Microsoft’s statistical MT engine uses two different decoders: A syntactically informed tree based decoder, which uses a parser building dependency treelets – better for translating between languages with different word orders. And a string-based decoder that needs no linguistic information to work. Good for fast training of language pairs without a parser.

Both make use of a set of statistical models. These models answer the question “how likely is B given A?”, and usually return an array of probabilities, which the decoder takes into account, giving each model a weight in the decision making process.

1.2 Continuous retraining
Microsoft’s MT system is retrained continuously, building fresh models for use in the decision making process. This is relevant for providing current
terminology and wide language coverage at any point in time.

The Microsoft Translator API includes methods and UI controls for collecting and submitting human edits of any human- or machine-produced translation suggestion, and allows humans to vote on these suggestions, where the vote of designated and authenticated humans is able to elevate the ranking of such an edit, so that it can be used transparently in subsequent translations using the standard basic Translate() API.

The service provides very simple methods:

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detect()</td>
<td>language detection</td>
</tr>
<tr>
<td>Translate()</td>
<td>translation</td>
</tr>
<tr>
<td>AddTranslation()</td>
<td>submitting an edit or a vote</td>
</tr>
<tr>
<td>GetTranslations()</td>
<td>retrieving a set of translations for a given source</td>
</tr>
</tbody>
</table>

Advanced methods include array functionality for all of the above. The service is also able to deliver voice format for the translations it produces.

The ratings and approvals are private to the user or enterprise who submitted it, an edit can appear as a suggestion elsewhere, immediately, and will be used in subsequent retraining of the service - underscoring the collaborative nature of human edits.

3 Collaborative Translations at Microsoft

Collaborative translations are being used at Microsoft in several places, of which we will describe two here.

3.1 Support Knowledge Base

The knowledge base uses a customized version of the Microsoft Translator engine. Any support personnel worldwide can visit the internal copy of the knowledge base and perform edits on any machine translated article using the web user interface. The web user interface mirrors the public view of the knowledge base, but with the addition of an edit option. The support personnel initiates an immediate republishing of the article. At that moment a simple automatic translation is triggered, using the human edit for any matching sentences, and the current MT answer for any other sentences.

Support personnel use this for any MTed article they discovered as being unclear, based on user contact. Only a very small portion of knowledge base articles are fully human translated.
3.2 Optimizing for the User

Machine Translating the Czech Knowledge Base

October 2009
- 2.5% of content of the English KB is human translated to Czech, ranked by page view.
- The top 2.5% cover an estimated 50% of the page views.
- The remaining content is untranslated.

January 2010
- 2.5% of content of the English KB is human translated to Czech, ranked by page view.
- The top 2.5% cover an estimated 50% of the page views.
- The remaining content is machine translated, starting December 5 and completed over the next 10 days.

Referrals from the Czech Republic
- Referrals to the CSS knowledge base site from the top 2 search engines in the Czech Republic (google.cz and seznam.cz
- to the Czech KB (cs)
- to the KB in other languages (All other languages)

3.3 Microsoft Developer Network

The Microsoft Developer Network web site allows users to submit edits of the machine translated content (in certain languages only).
Any user can submit edits. The enthusiasts who have shown capable of submitting high quality edits are elevated to trusted status, and their edits are being approved and visible to everyone immediately. For content and locales that does not have sufficient approvers, LSPs are tasked with the job.

4 Data and Quality

Collecting human edits is just one way of growing the supply of training data for a language and domain. Other methods can and should be applied as well. The following lists the effects of growing training data for use in an SMT system.

4.1 Data Sources

Collaborative Translation Store
Selected Web Content
- After page and sentence alignment
Existing (mostly) parallel data
- Microsoft manuals
- Dictionaries, phrasebooks
- Government Data
Data sharing associations
- Linguistic Data Consortium,
- Taus Data Association,
- ELRA,
- Others
Licensed data
- Microsoft Press, …
Comparable (non-parallel) data
- Wikipedia
- News articles
Number of parallel sentences

Human evaluation results

4.2 Adding Domain Specificity

Custom Target language models and phrase- or treelet tables serve to enhance adherence to the intended style and terminology.

Example: Chinese for Sybase (BLEU)

<table>
<thead>
<tr>
<th>System Description</th>
<th>General</th>
<th>Microsoft</th>
<th>Sybase</th>
</tr>
</thead>
<tbody>
<tr>
<td>General domain</td>
<td>14.26</td>
<td>29.74</td>
<td>34.81</td>
</tr>
<tr>
<td>Microsoft</td>
<td>12.32</td>
<td>34.65</td>
<td>29.95</td>
</tr>
<tr>
<td>Microsoft with Sybase</td>
<td>12.16</td>
<td>34.66</td>
<td>30.24</td>
</tr>
<tr>
<td>General and Microsoft</td>
<td>15.38</td>
<td>35.80</td>
<td>44.49</td>
</tr>
<tr>
<td>and TAUS</td>
<td>12.57</td>
<td>29.51</td>
<td>47.16</td>
</tr>
</tbody>
</table>

Example: German for Sybase (BLEU)

<table>
<thead>
<tr>
<th>System Description</th>
<th>General</th>
<th>Microsoft</th>
<th>Sybase</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Domain</td>
<td>25.19</td>
<td>40.61</td>
<td>34.85</td>
</tr>
<tr>
<td>Microsoft</td>
<td>21.95</td>
<td>52.39</td>
<td>41.55</td>
</tr>
<tr>
<td>Microsoft with Sybase</td>
<td>22.83</td>
<td>52.07</td>
<td>42.07</td>
</tr>
<tr>
<td>General and Microsoft</td>
<td>23.86</td>
<td>52.72</td>
<td>48.83</td>
</tr>
<tr>
<td>and TAUS</td>
<td>19.44</td>
<td>37.27</td>
<td>50.85</td>
</tr>
</tbody>
</table>
References


Microsoft Translator: www.microsofttranslator.com

TAUS Data Association: www.tausdata.org