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.

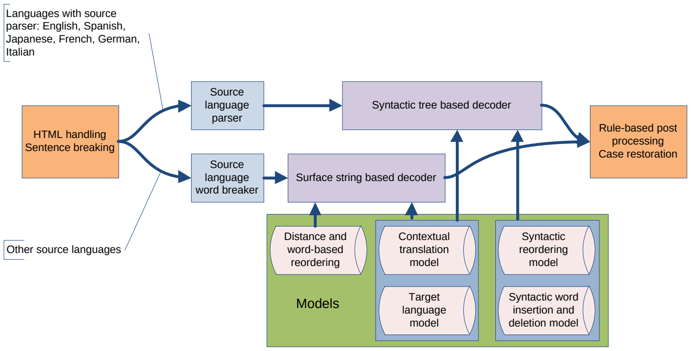
1 Introduction

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

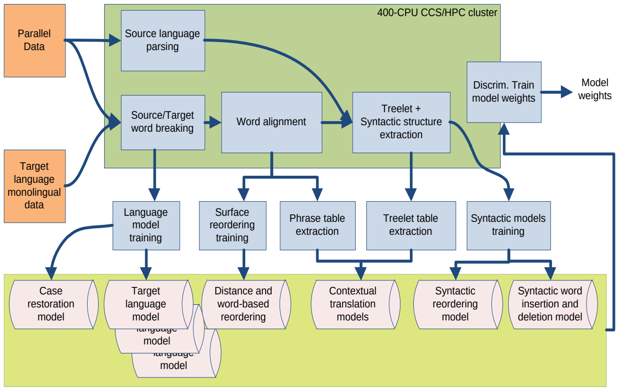
Microsoft's statistical MT engine uses two different decoders: A syntactically informed tree based decoder, which uses a parser building dependency treelets – better for translating between languages with different word orders. And a string-based decoder that needs no linguistic information to work. 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.


Languages with source parser: English, Spanish, Japanese, French, German, Italian
HTML handling Sentence breaking
Source language parser
Syntactic tree based decoder
Rule-based post processing Case restoration
Source language word breaker
Surface string based decoder
Other source languages
Distance and word-based reordering
Contextual translation model
Syntactic reordering model
Models
Target language model
Syntactic word insertion and deletion model

1.2 Continuous retraining

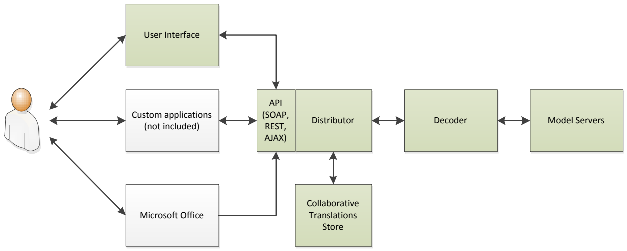
Microsoft's MT system is retrained continuously, building fresh models for use in the decision making process. This is relevant for providing current terminology and wide language coverage at any point in time.


400-CPU CCS/HPC cluster
Parallel Data
Target language monolingual data
Source language parsing
Source/Target word breaking
Word alignment
Treelet + Syntactic structure extraction
Discrim. Train model weights
Model weights
Language model training
Surface reordering training
Phrase table extraction
Treelet table extraction
Syntactic models training
Case restoration model
Target language model
language model
language model
Distance and word-based reordering
Contextual translation models
Syntactic reordering model
Syntactic word insertion and deletion model

2 Collaborative Translations

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 submissions, edits and ratings are stored online, and used as an integral part of the MT service itself.



The service provides very simple methods:

Detect()	language detection
Translate()	translation
AddTranslation()	submitting an edit or a vote
GetTranslations()	retrieving a set of translations for a given source

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

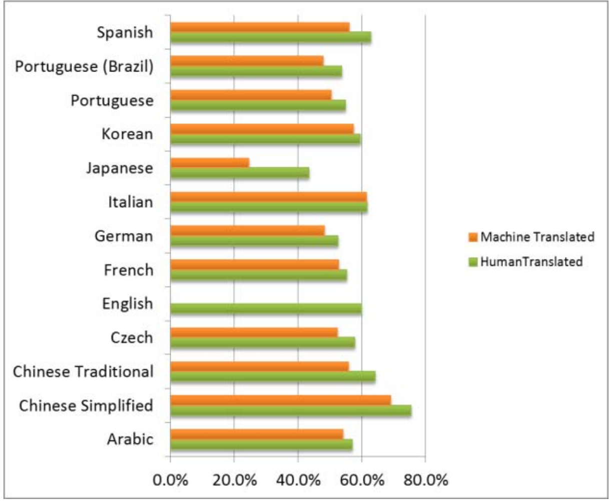
Collaborative translations are being used at Microsoft in several places, of which we will describe two here.

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.

Knowledge Base Article – Success rate:



3.2 Optimizing for the User

Machine Translating the Czech Knowledge Base

October 2009

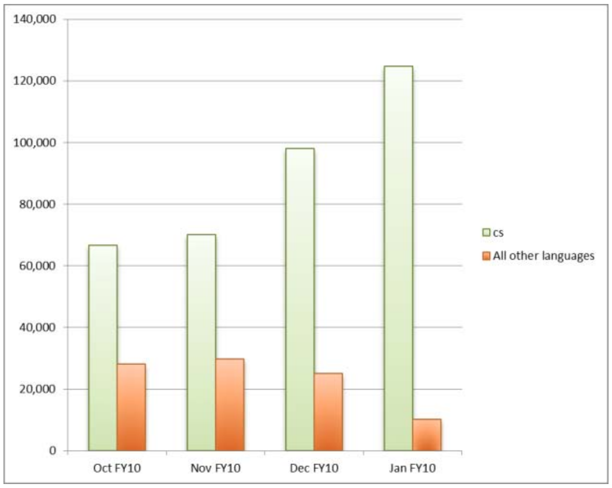
- 2.5% of content of the English KB is human translated to Czech, ranked by page view.
- The top 2.5% cover an estimated 50% of the page views.
- The remaining content is **untranslated**.

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.

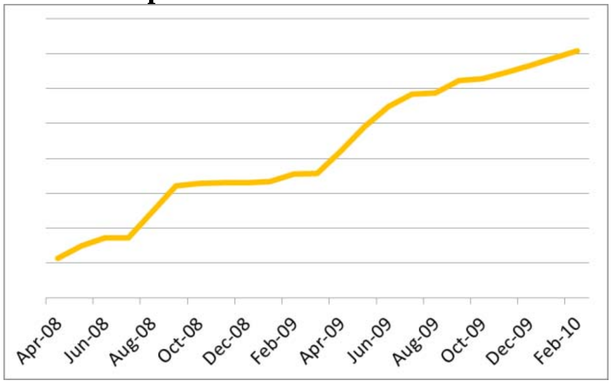
4 Data and Quality

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.

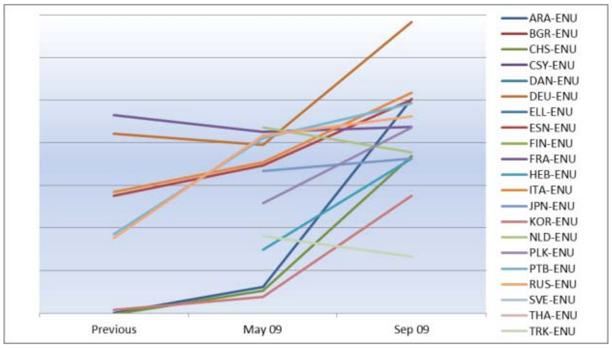
4.1 Data Sources

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

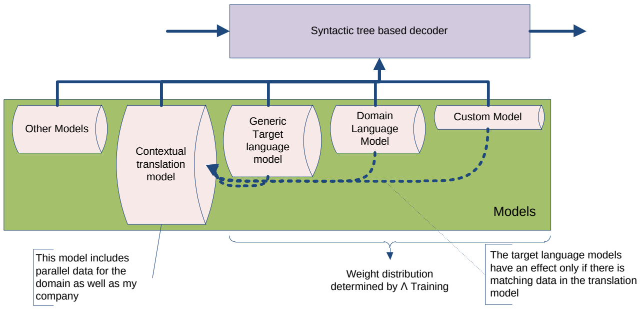


Human evaluation results



4.2 Adding Domain Specificity

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


Syntactic tree based decoder
Other Models
Contextual translation model
Generic Target language model
Domain Language Model
Custom Model
Models
This model includes parallel data for the domain as well as my company
Weight distribution determined by Λ Training
The target language models have an effect only if there is matching data in the translation model

Example: Chinese for Sybase (BLEU)

	Test Set		
System Description	General	Microsoft	Sybase
General domain	14.26	29.74	34.81
Microsoft	12.32	34.65	29.95
Microsoft with Sybase	12.16	34.66	30.24
General and Microsoft and TAUS	15.38	35.80	44.49
Above with Sybase lambda	12.57	29.51	47.16

Example: German for Sybase (BLEU)

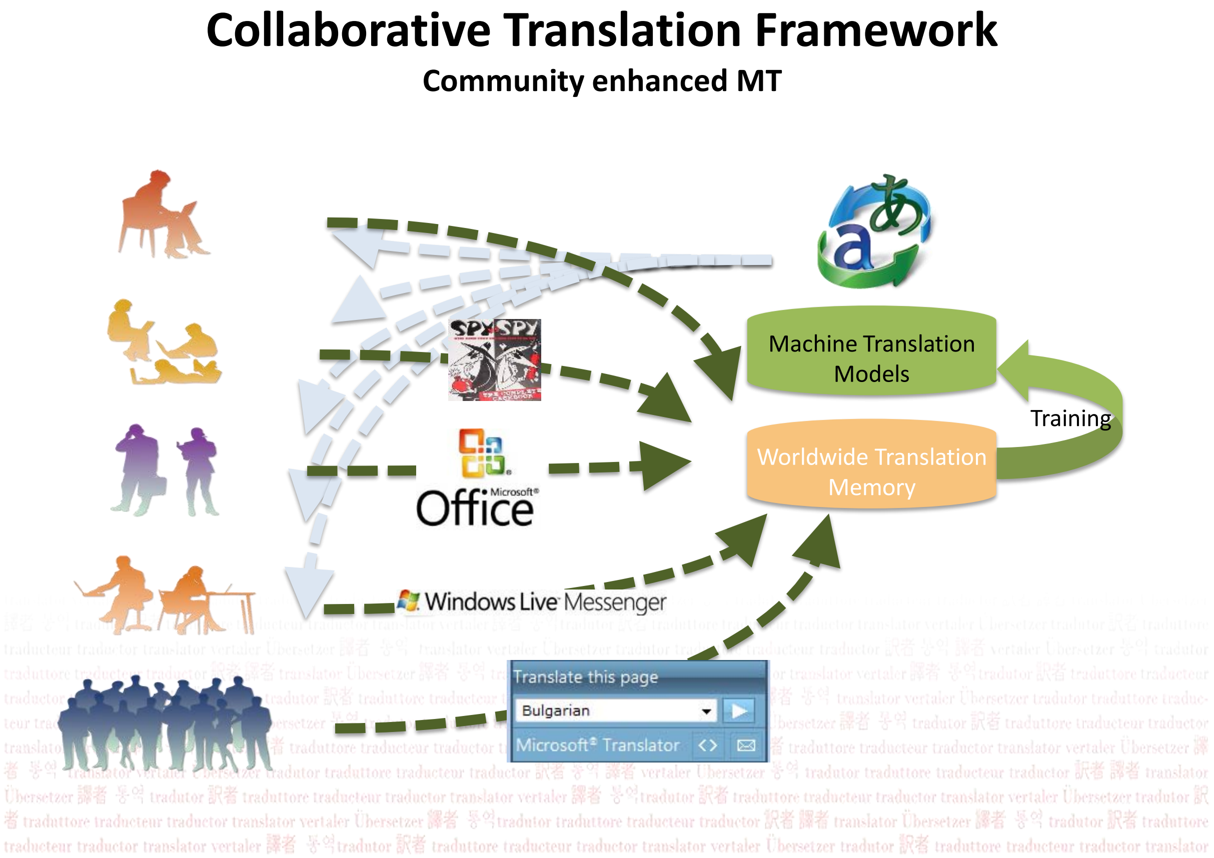
System Description	General	Microsoft	Sybase
General Domain	25.19	40.61	34.85
Microsoft	21.95	52.39	41.55
Microsoft with Sybase	22.83	52.07	42.07
General and Microsoft and TAUS	23.86	52.72	48.83
Above with Sybase lambda	19.44	37.27	50.85

References

Chris Quirk, Arul Menezes, and Colin Cherry: *Dependency Treelet Translation: Syntactically Informed Phrasal SMT*, in Proceedings of ACL, Association for Computational Linguistics, June 2005

Microsoft Translator: *www.microsofttranslator.com*

TAUS Data Association: *www.tausdata.org*



Collaborative Translation Framework

Community enhanced MT