USE CASE: Customization and Collaboration to enhance MT for a Knowledge Base online portal

Chris Wendt
Microsoft Research
Redmond, WA USA
Chris.Wendt@microsoft.com

Federico Garcea
Microsoft Research
Redmond, WA USA
Federico.Garcea@microsoft.com

Abstract

The Microsoft Customer Support organization maintains a large Knowledge Base online portal, which contains over 200,000 active documents and provides localized versions in more than 40 languages. The portal is a large-scale example of an effective use of various advanced techniques of human and machine translation working together. We describe how the Knowledge Base site has evolved from an established raw MT and post-editing scheme to the current combination of human and machine translation, continuing to report high rate of customer satisfaction and effectively engage satellite offices and IT experts around the world to maintain an updated and high quality information base.

1 Introduction

The use of machine translation (MT) coupled with post-editing by professional translators has become a common technique for localization of IT documentation, online and offline. In recent years, we have also seen the emergence of crowdsourcing translation (Wendt, 2012; Ficcarelli, Litsl, and Vahldieck, 2012), while the field of MT research has progressed in the areas of domain adaptation (Moore and Lewis, 2010) and of customization (Lewis and Yang, 2012).

While each approach has its merits and improves the quality of a basic general MT approach, there is additional value in combining several approaches and making them available in the same platform or ecosystem. One can devise an iterative workflow that uses a combination of adapatation, customization, collaboration, and create a system that will improve satisfaction of end users, engagement of top contributors in a crowd, and finally it can improve and inform the quality of an MT system.

2 Knowledge Base Portal publication workflow

Localization of technical documents for IT domain has been a solid use case for application of post-edited MT and even raw MT, leading to higher end user satisfaction (Smets and Riesco, 2012).

In fact, Microsoft Research's first application of MT was specifically for localization of technical documentation.

The Knowledge Base portal contains frequently updated information on known issues on a large array of software products and platforms, including security bulletins. Millions of IT professionals and software users visit it daily, and it is connected to user-supported forums as well as product information and marketing material.

The Customer Support Services team at Microsoft (CSS) maintains and manage the portal and its content.

2.1 Publication of source content

New content is published weekly, and the original articles are always written in English. Selected documents are professionally translated in selected markets and published as soon as they become available.

2.2 Automatic Translation

As part of the publishing workflow, a Machine Translation Widget is applied to newly published documents in English.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 401–405.

The translation widget uses a domain-adapted MT system, if available for the target language, and a Collaborative Translation Memory to match previously edited segments (sentences).

2.3 Publication of translated content

The resulting translated documents are published on the localized sites, so search engines can index them and provide a permanent link for social media sites and reference from other online documents or emails.

2.4 Collaborative Post-Editing

The CSS team maintains a list of 'community leaders' or most-valuable players (MVP) that are actively engaged in improving quality of the published content in their language.

The MVPs are notified when new content is available, and they can review and edit the published content at their convenience.

2.5 MT Customization

Documents in the same domain that have been professionally translated and community edits are used periodically to train and evaluate customized MT systems, which are then deployed for online use and consumed immediately by user requests for localized content.

3 Machine Translation: domain adaptation and customization

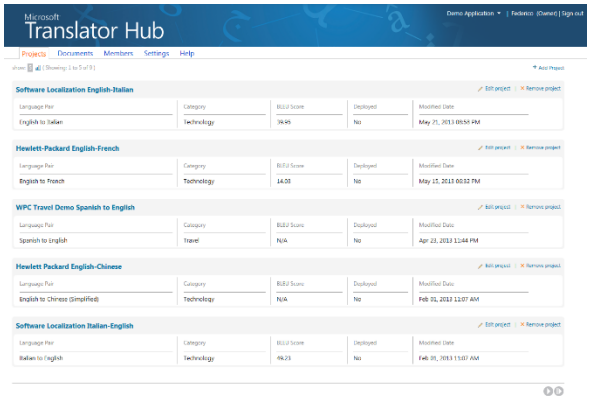
Supporting internal localization teams through MT has always been a primary application of Microsoft Research MT technology. Based on recent advances in domain adaptation techniques (Moore and Lewis, 2010), the MSR team has trained and deployed a set of translation systems for the IT and software domain (also referred to just as the *technology* domain), covering more than 30 different markets.

The quality of domain-adapted systems is between 2 and over 10 BLEU points compared to the general domain systems.

The use of these translation systems has proven effective with helping end users. With the exception of Korean and Japanese markets, in all other cases, user satisfaction is within 5% between Human Translated documents and Machine Translated (*raw MT*) documents (Smets and Riesco, 2012).

In 2012, Microsoft released the Microsoft Translator Hub, which allows any third party to create their own SMT system, including managing online deployments and usage.

With the availability of Hub, organizations can submit parallel and monolingual data in domain and train their customized and domain-specific translation system. The models that Microsoft Translator already utilizes in the publicly available translation systems (translation models, language models, order models, etc.) can be optionally included, so the resulting translation system is a domain-adapted system that has learned from both existing large models for general domain and models created from customer-supplied data.



The Hub systems produced by the CSS organization for KB site are in essence very similar to what Microsoft Research has produced for their consumption in the past. With a self-serve model, the turnaround to make new language pairs available or to retrain existing language pairs with different data is much shorter, i.e. it is a matter of days rather than months.

Because the CSS organization has already access to content previously localized or translated, it has been easy for them to grow the number of supported language pairs.

Within a few weeks, they were able to train and publish nine additional systems. Each one was trained with around 100,000 parallel sentences, and each one showed an increase in BLEU score against their own in-domain test set compared to the publicly available systems.

As seen in the table below, the increases are all in multiple percentage points.

Target Language	Sentence Pairs	BLEU	Baseline BLEU	Diff from baseline
Estonian	103,842	43.43	38.74	4.69
Latvian	98,089	56.17	38.88	17.29
Slovak	127,140	53.61	47.35	6.27
Ukrainian	111,677	43.74	37.38	6.36

Even in a narrow domain like KB articles for Microsoft software, MT still provides a *gisting* translation and not a fluent translation in many language pairs and in many cases. In order to improve quality and fluency of the translations, customized MT can be paired with other techniques.

4 Handling of Terminology

Preserving consistent terminology in the translation of technical documents has been a challenge for Machine Translation systems.

A customized MT that has been trained on parallel documents and target language monolingual documents in-domain can mitigate the terminology problem, but it doesn't completely resolve it.

In 2013, Microsoft Translator Hub introduced a Dictionary feature, to let customers import a dictionary of terms (phrases) with their desired translations. The resulting MT system will always respect the dictionary translations, i.e. it will always favor a dictionary entry over any statistical hypothesis from the models.

While this is expedient and effective in a variety of cases, it can also be problematic, since it may alter fluency of the whole sentence by forcing an incorrect translation, e.g. it may produce concordance errors and other grammatical errors.

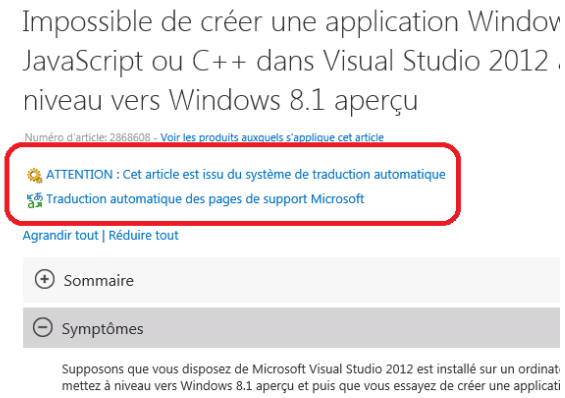
Another approach to mitigate the terminology problem is to make corrections using the Collaborative Translation Framework features (described more in detail in the next section). Since corrections apply to a whole sentence, they are less effective and require lot of effort, though it may pay if content published is usually repetitive in nature.

Ideally, a good terminology solution could inform an MT system and be part of a fast, incremental training, and be context-aware – with the context being a single document or an entire collection or site. There has been interesting prior work on this topic (Hardt and Elming, 2010), and it is certainly an area for further research and development.

5 Collaboration

The KB site display pages by default in the end-user browser locale. Certain pages are professionally translated, but most contain MT content.

All pages that contain MT content bear a special icon and a link to a disclaimer on top, which is a good established practice (see fig. below)



Following the disclaimer link takes users to a page that describes how MT works, and how to help improve its quality.

In particular, it contains guidelines on what to pay attention for in reviewing MT content and what to correct. For example:

- Words are dropped: sometimes, negation is dropped in the MT version;
- Command lines are corrupted by MT
- Terminology errors: for example, " driver" translated as the conductor of a vehicle
- Phrases left in English when they should be translated

It also provides guidelines on what to leave alone: grammatical errors or word order errors do not need to be corrected, unless they make the meaning difficult to understand. Indeed, it is not possible to correct everything, because there are too many possible issues and too many articles.

Anybody can review and submit corrections, anonymously or using their own Microsoft Account (formerly LiveId).

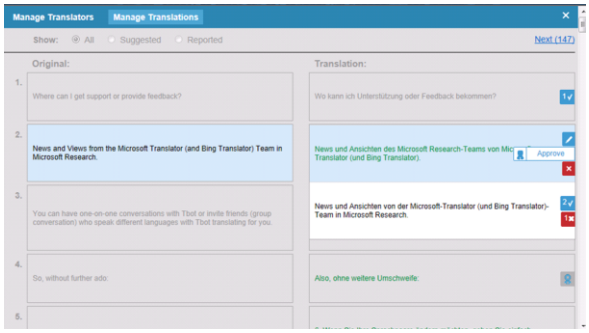
With Microsoft Translator CTF Widget, the administrators have access to a dashboard where they can nominate and invite users and give them the role of Reviewer: a reviewer has higher au-

thority than normal users, and higher authority or rank than MT system. This means that a correction submitted or 'promoted' (voted) by a reviewer or administrator will become the default translation for that sentence on the site, replacing the MT output.

Because the CSS team has an existing relationship with more than 500 MVPs around the world and hundreds of other IT professionals that contribute to their user forums and support sites, they already have access to a network of trusted reviewers.

This ongoing engagement produces thousands of weekly edits, constantly improving the quality of the information provided, especially for most frequently read and most important bulletins.

An existing system of rewards (points) for community contributors is also extended to translation editors and contributors.



Reviewers have access to the CTF dashboard (fig. above): here one can quickly see at a glance all 'pending' suggestions, i.e. suggestions and edits made by non-authoritative or 'normal' users: a reviewer can quickly scan and approve or reject tens or hundreds of translations.

This is somewhat similar to what Autodesk implemented in their Translate-It platform (Ficcarelli, Listl, and Vahldieck, 2012), which proved effective in a few pilot products.

All corrections together form a Translation Memory (TM) hosted in the cloud and accessible from anywhere.

As content from KB article consistently uses the same language and contains a number of boilerplate information and sentences, having this cloud TM proves effective in improving rapidly quality in a high number of pages.

Contributed translations have a utility beyond the immediate improvement of readability: they form an ever-growing corpus of parallel sentences in the same domain that can be used to train a customized system.

In fact, if the same administrative account is used for both accessing Hub and managing a CTF Widget on a site, it is straightforward to review, import, and reuse community edits to train a customized system. That is what the CSS team is doing to expand the set of available language pairs, using the information collected via the KB site in conjunction with past professionally translated content to build more customized systems.

In addition, the CSS team also employs professional translators to do post-editing work on the site using the same widget. This is particularly useful for new content that may regard brand new products and offering, hence documents that likely contain new terminology and brand names that were previously unknown to both MT and TM systems.

This accelerates again the learning cycle for both MT engines – which can get now easily retrained and re-deployed within two days – and for human editors, who can immediately see new contributions by other experts and learn how to consistently edit with new terminology.

6 Conclusions

We have described how the KB portal maintained by the CSS team at Microsoft makes use of a combination of technologies to bring human-augmented machine translation to fruition of its millions of users worldwide.

The goal of this approach is to improve user satisfaction while maintaining the cost of localization constant, and to continue to expand the number of markets and end users that are covered by the documentation available.

Let us summarize the technologies used to maintain this translation workflow:

- A content management system for the original documents in English and professional translations.
- A customized CTF Widget that provides a desired custom user experience over the standard widget provided by Microsoft Translator.
- A Microsoft Translator Hub application space to train and manage customized MT systems.
- A portal to manage statistics and analytics of contributions and to manage volunteers, freelance translators, and to assign points or rewards to contributors.

The whole system can be administered and managed by localization project managers and

experts, and does not need MT experts, though it needs PMs that have experience with MT and its dissemination.

We believe the experience that CSS has built over the years is a great example of how to leverage the best technologies for both Machine and Human translation. The workflow is distributed and decentralized, and it easily lends itself to long distance collaboration and crowd sourcing.

CSS evaluates periodically impact on users, and we look forward to present latest findings once more new language pairs that have been self-trained and self-managed are available and tracked by their surveys.

7 Future Directions

There are a few areas where additional research and design work can improve the state-of-the-art, in particular:

- Traditional analytics for a web site that measure page visits can be used to solicit corrections, and could be coupled with NLP tools to provide automated elicitation, e.g. extract key sentences and terms that need to be translated or reviewed.
- Incremental online training of translation models from user edits.
- Morphology and context-aware terminology handling.
- CTF Plugin or support for crowdsourcing platforms, to enable and engage existing crowds to provide corrections.
- Automated retrain and deployment of customized MT systems built from user contributions.

References

Amittai Axelrod, QingJun Li, and Will Lewis, 2012. Applications of Data Selection via Cross-Entropy Difference for Real-World Statistical Machine Translation, *in Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2012), International Workshop on Spoken Language Translation (IWSLT)*.

William D. Lewis and Phong Yang, 2012. Building MT for a Severely Under-Resourced Language: White Hmong, *Association for Machine Translation in the Americas.*

Robert C. Moore and William Lewis, 2010. Intelligent Selection of Language Model Training Data, *in Proceedings of the ACL 2010 Conference Short Papers, Association for Computational Linguistics,* Uppsala, Sweden.

Daniel Hardt and Jakob Elming, 2010. Incremental Re-training for Post-editing SMT, *Ninth Conference of the Association for Machine Translation in the Americas 2010,* Denver, USA.

Martine Smets and Jose Luis Riesco, 2012. Continuous Publishing and Localization*, TAUS 2012*, Paris

Chris Wendt, 2012. Collaborative Machine Translation: You, Your community and Microsoft's knowledge working together. *TAUS Asia Translation Summit,* Beijing.

Simona Ficcarelli, Silvia Listl, and Bodo Vahldieck, 2012. The Autodesk Collaborative Translation Platform, *Localization World*, Paris.