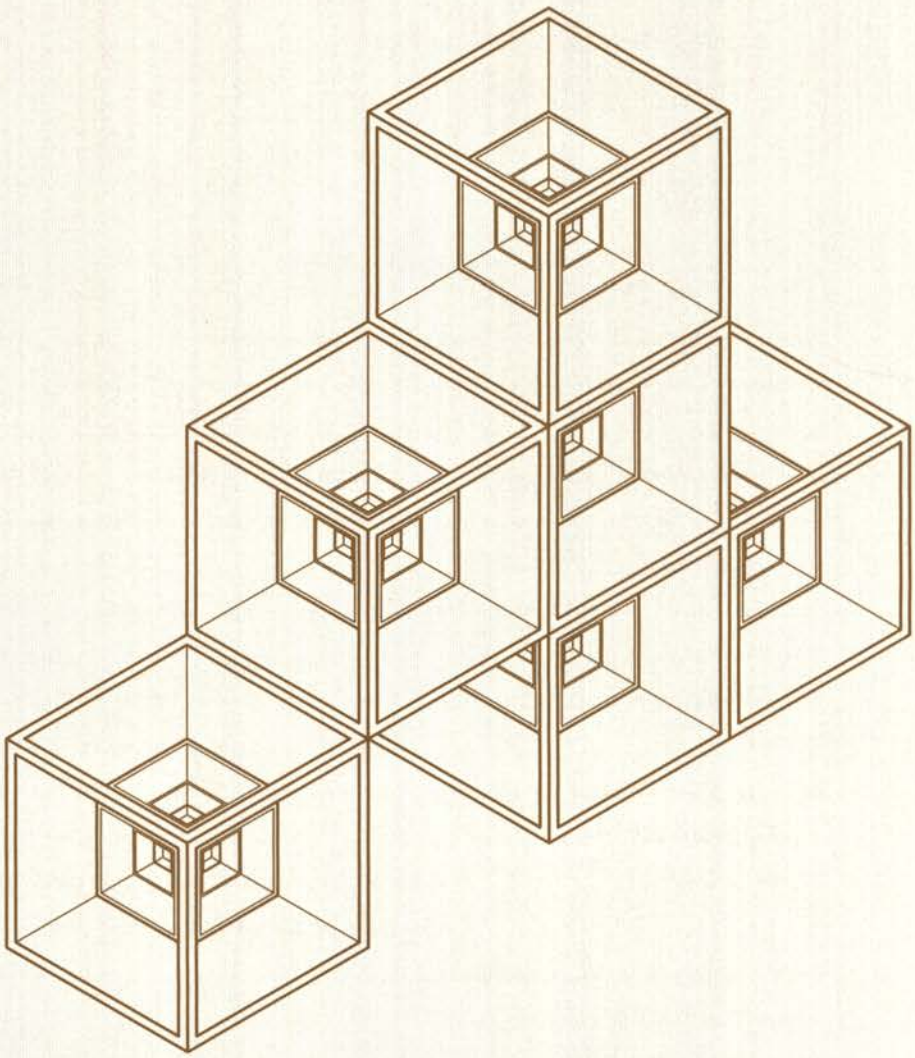
ISSN 1883-1818

AAMT Journal

Asia-Pacific Association for Machine Translation



September 2012 No.52

アジア太平洋機械翻訳協会

目　次

CONTENT

Foreword

Solving the Translation Industry Dilemma

In five self-help steps

AUTHOR: JAAP VAN DER MEER, DIRECTOR AND FOUNDER OF THE TRANSLATION AUTOMATION USER SOCIETY (TAUS)

Communication across the world's many spoken languages is a problem. The technology that can help solve this problem is getting better and better. But the professionals, needed to use and improve the technology – at least the vast majority of them – reject the technology and deny its advancements because they fear for their jobs. A good Dutch expression jumps to mind: *"Man suffers most from the suffering he fears, but never appears."*

In this prefatory note I argue that to develop the MT industry further we need to push users to overcome their fears – MT technology is already quite good in its current state – and become supporters and enthusiastic users. I am proposing "five self-help steps" with links to the TAUS web site for more support to users who are ready to adopt MT technology in their daily practice.

The Translation Industry Dilemma

The translation industry is caught in a dilemma: *to automate or not to automate*. Dammed if you do, dammed if you don't. Using machine translation technology feels like a curse to everyone who spent years to study other languages and become a professional translator. But if you don't, it's becoming harder to stay in business. Customers want translations faster and cheaper. Besides, the volume of information requiring translation is beyond imagination. The worldwide population of a few hundred thousand professional translators is just scratching the surface. Machine translation technology can help increase efficiency and open new business opportunities.

TAUS recommends five steps for every buyer or provider of translation who is caught in this dilemma.

1. Be rational

The first step is to 'get real' about the technology, ask the right questions. Since MT has been made available by Google and Microsoft and millions of people started using it every day, have we seen a decline in the mainstream translation market? Have translation jobs been taken over by computers? Market analysts report solid growth rates of the translation industry from 2005 to 2011. The need for translation is so great; it is becoming impossible to meet demand without the use of automated translation technology. Did other industries decline or suffer when automation technologies were introduced? Think of the banking industry or the travel industry. Admittedly jobs changed, but overall opportunities have been on the rise and industries started to flourish when new technologies were introduced. The first step is therefore to open up, get over the fears for change. TAUS has published a series of articles and white papers about changes in the translation industry and

language business innovation. We welcome you to surf through this library of publications (http://www.translationautomation.com/articles/)

2. Try it out

The second step is to try it out. After all, the proof of the pudding is in the eating. However, don't start trying MT before you have successfully concluded step 1. Too many people have already trialed MT with the sole intention to prove that it isn't working, that it's ugly and often laughable. You want to try it looking at the potential and the positive. TAUS recommends a simple do-it-yourself test before engaging vendors and consultants. There are many options available to try machine translation technology. You can check out the TAUS Tracker (http://www.taustracker.com) online directory of MT engines to see what is available in your preferred language pair. You can translate some of your own text using online free MT tools, but you will in most cases not be able to feed these engines with your own terms and phrases to help them improve the results. It is worth buying a desk-top license of a commercial MT package. For a small investment (usually under 1,000 Euro) you get access to a rich set of customization features, often not much different from the features offered with the server-based systems. If your team is equipped with some software engineering skills and you are prepared to set aside some time, the other option to gain MT experience is to download open-source MT software, such as the Moses engine. Moses is gaining popularity very quickly now among service providers and buyers of translation. TAUS has published a (free) online Moses tutorial (http://www.tauslabs.com/open-source-mt/mosescore/moses-tutorial) , dozens of use cases and best practice reports (http://www.translationautomation.com/reports/) for training open-source MT systems.

3. Learn from others

Now you are on your way to become a user of MT, you know you cannot avoid going through trials and errors. The third step we recommend, to not make too many of the mistakes others made, is to learn from others. TAUS has launched an online knowledge base for MT users: the FAQ Forum (http://www.tauslabs.com/open-source-mt/faqs) Questions from users are posted, intelligence is collected from all other users and TAUS is undertaking further research. Responses are documented and reviewed. Another source of intelligence is the TAUS YouTube channel (http://www.youtube.com/user/TAUSvideos) with presentations of use cases. Finally if you can't find what you are looking for you can always send an email to info@translationautomation.com. Only after you have completed your discovery of others' trials and errors, we recommend that you start documenting your goals and addressing the fundamental questions about budget, vendors and business models.

4. Measure and benchmark

If, after you have learned your lessons, you decide to proceed and implement MT technology in your organization, the most crucial step (of all five) is to define your goals. Which content types do you plan to apply the technology to and what do you want to achieve in terms of quality and productivity. Many large and small enterprises struggle to formulate these goals, not just for the use of MT, but even for human-based translation. The lack of criteria and clear evaluation metrics leads to uncertainty, friction, disputes, and loss of time and money. It is often opinions, rather than measurements, that lead to the rejection of MT system and the firing of vendors or translators. In consultation with many of the large enterprise members TAUS has developed the Dynamic Quality

Framework (DQF) (http://www.tauslabs.com/dynamic-quality/about-dqf). DQF allows users to profile their content types, based on a measurement of three factors: utility, timeliness and sentiment (UTS). The resulting UTS score is then linked to a recommended approach to evaluating the quality of the translation of this particular content. DQF is a knowledge base documenting seven different quality evaluation approaches. In addition DQF provides industry-shared tools that allow users to compare and benchmark MT productivity and translation quality scores, such as adequacy and fluency. The TAUS Dynamic Quality Framework brings credibility and transparency to translation business. Without the methods and the knowledge to set goals, to measure and to benchmark against these goals, users of machine translation technology are like hunted game.

5. Share and collaborate

The fifth and final step that TAUS recommends is strategic and is aimed at sustaining growth and innovation in the translation industry. Just like you have learned from others at step 3, we recommend that you share the lessons you have learned, so that others can learn from you. Do not fear to give up an advantage, or you will soon find that it's you who is left behind. Only if we decide to be truly open and ensure complete interoperability in the systems we build and use, we will reach the maximum growth opportunities of the translation industry. TAUS has launched a translation web services API (http://tauslabs.com/interoperability/taus-translation-api) to facilitate seamless exchange of translation jobs on the web. But we take sharing even a step further. TAUS advocates that all organizations share their language data in order to improve the performance of their own translation technologies and to stimulate innovation and growth in the translation industry overall. More and more companies and government organizations are following this vision. In 2008 TAUS launched a data sharing platform. Today the TAUS Data repository (http://www.tausdata.org/) contains already 50 Billion words of shared translation memories in more than 2000 language pairs. These language data are available for every user of MT to train and improve their engines.

追悼　白井諭さん

真摯な言語分析に基づく言語処理を貫いて

日本電信電話株式会社／アジア太平洋機械翻訳協会会長
中岩　浩巳

私の元上司であり、元同僚であり、ご指導を仰いだ、白井諭さんが、2012年1月19日に逝去されました。56歳という若さで、突然のことでした。本当に残念です。

白井さんには私がNTT研究所に入所した際、直接指導していただきました。学生時代は音声信号処理の研究をしていた私が、NTT入社後、自然言語処理という今まで経験していなかった分野の研究室に配属になったのですが、言語現象の面白さ、奥の深さを熱く語ってくださり、私を自然言語処理の研究者に導いてくださいました。その時に、与えていただいた研究テーマの一つが日英機械翻訳における省略補完技術でした。日本語では主語や目的語が頻繁に省略されるのに対し、英語では明示的に訳す必要があります。このため、日本語解析の段階で省略された個所を補って翻訳する技術が必要となります。日本語には必須となる要素が明示的ではないため、日本語の省略補完を考える際に、必須となる要素はなにかの定義が難しく、場合によってはその判定が恣意的になりますが、日英機械翻訳というタスクでは英語で訳す必要のある要素という明確な限定が出来るため、必須性判定が明確になります。その後、私がこの研究を続けることが出来たのは、白井さんの的確な課題設定と、ご指導の賜物であると感謝しております。

白井さんは、NTT入社以来様々なご研究をなさってきましたが、私が印象に残っているのは、日本語の係り受け解析と、日英機械翻訳のための日本語事前書き換え技術です。係り受け解析では、日本語の表現を言語学者時枝誠記、三浦つとむが提唱した言語化定説に基づく日本語文法をベースに言語現象を事細かに分析され、係り受け解析規則を体系的に構築されました。このシステムはNTT研究所が開発したルールベース型日英機械翻訳システムALT-J/Eの中に組み込まれ、同システムの翻訳性能向上に貢献しました。また、日本語事前書き換えは、日本語と英語の語順や品詞の違いを克服するために、日英変換を行う前の日本語係り受け解析後の段階で、英語に翻訳しやすい語順、表現からなる係り受け構造に書き換えるというものです。この考え方は、最近研究の主流となっている統計的機械翻訳でも採用されている先駆的なものであり、白井さんの日英翻訳の本質を見抜く力と、先見性には目を見張るものがあり、心よりの敬意を表したいと思います。このような白井さんのご研究も含めたNTT研究所での日英機械翻訳の成果が、内外から注目され、日本語語彙大系として岩波書店から出版され、世に出すことが出来ました。白井さんも編者として、専門性が高く通常の辞書とは大幅に異なる同書を、広く受け入れられるものにすべく、熱心に議論されていたのが昨日のように思い出されます。

また、白井さんは、ATR音声翻訳研究所に機械翻訳の研究室の室長としてもご活躍され、音声翻訳の分野でも多くの研究者を指導されるとともに、ご自身も優れた研究成果を生みだされました。後任として同研究室の室長に着任した際には、ATRでのプロジェクトの進め方などについてご指導いただきました。先日の告別式には、当時のATR関係者が多数参列され、白井さんの人望の厚さが伺えました。

その後、白井さんは、NTTアドバンステクノロジ社に移られ、言語処理以外の仕事も担当されるようになりましたが、言語処理の仕事を自ら提案し、言語処理の仕事に携わろうとされる努力は精神的にも肉体的にも大変な労力がいることだったと思い

ます。白井さんの言語現象の分析に対する情熱は、ずっと保たれたままでした。

ちょうどそのころ、前出の ALT-J/E の開発リーダであり、その後鳥取大に移られた故池原悟先生とは、「言語・認識・表現」LACE 研究会の創設に貢献され、同研究会の年次大会では、白井さんの研究に対する考えを発表され、熱い議論をされていたのが記憶に残っております。また、池原先生が中心となった CREST のプロジェクト「セマンティック・タイポロジーによる言語の等価変換と生成技術」にも、白井さんは CREST プロジェクトの研究者として、また、池原先生の右腕として参加され、ALT-J/E ではカバーできていなかった重文複文構造の大規模変換パターンの構築に貢献されました。ほぼ月に 1 回行われていたミーティングでの、白井さんからの本質的かつ的確なコメントが印象に残っております。

白井さんは、現実の言語現象に向き合い、その本質を見ようと追求する姿勢に妥協がありませんでした。気になる言語現象について質問すると、ご自分の考えを次から次へと様々な実例を用いて、喜々として説明して下さり、普段から、言語現象に注意していらっしゃることがうかがえました。その現れとして、特異な言語現象を発見すると、コーパスとして蓄積し続けており、その質の高さは、他の研究者の方が舌を巻くほどでした。白井さんが収集されたコーパスの質の高さをご存知の方は、そのコーパスが白井さんの逝去で散逸するのは、言語処理界の損失と嘆かれる方もいらっしゃるほどです。

個人的には、20 年ほど前になりますが、鳥取で行われた学会に参加した際に、白井さんと二人で鳥取市内にほど近い温泉に宿泊し、夜遅くまで少しお酒を飲みながら研究の話をしたのが昨日のように思われます。

以前は単身赴任をされておりましたが、最近はご自宅に戻られ、ご家族との楽しい生活をされていたのに突然の出来事で、奥さま、ご子息には、おかける言葉がありませんでした。ご家庭での白井さんは、会社や学会でお話しする白井さんとはまた違った面をお持ちだったとお聞きしております。

ご冥福を心からお祈りいたします。

Project Report

Machine Translation Deployment For Cisco Technical Documentation

By Dieu Tran

Cisco G.K., Tokyo, Japan,

Business Operations Manager, APJC Region, Gobal Shared Services Organization

1. Machine Translation Opportunity

Japan is a very important market for Cisco. To make it easier for its Japanese partners and customers to do business with Cisco, the company is investing in a web self-service model where the most current technical documentation (*Tech Docs*) is regularly translated and published on the company's enterprise website (Cisco.com). By doing so, Cisco can also maintain its competitiveness in the local market, since technical manuals are often used as supporting material for product evaluations (as part of purchase decisions), deployment, operation and application development. Given the high volume of the technical library – an estimated 2 billion words spanning a portfolio of over 1,000 product series across multiple networking architectures – Cisco has put in place a phased approach that includes Machine Translation (MT) technology to effectively scale the localization delivery model to address the need to respond to market demands for more localized contents (see Figure 1.1).

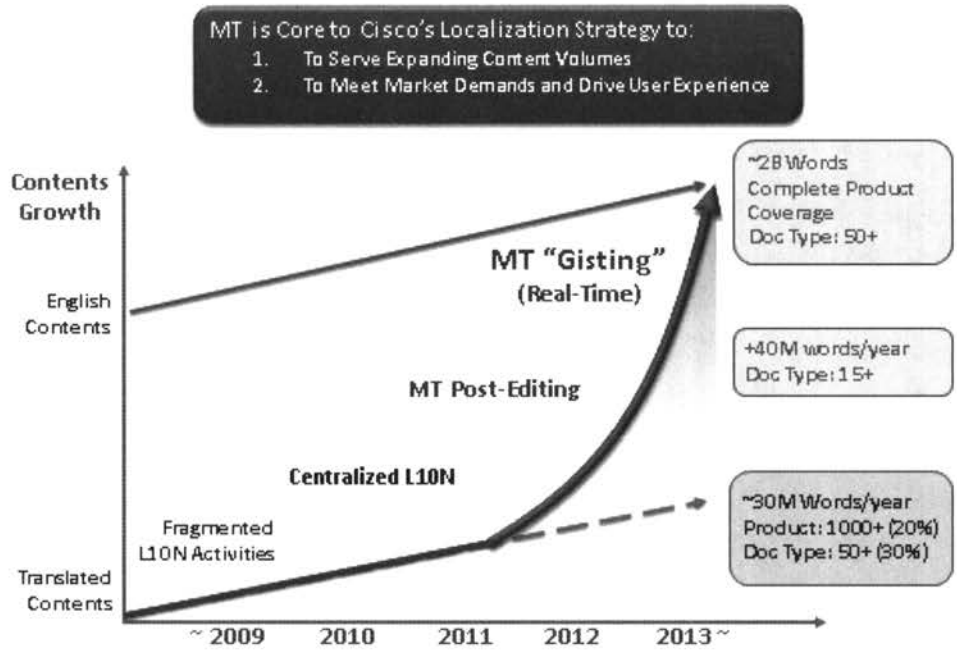


Figure 1.1.: MT rollout approach for TechDoc

First, all existing localization efforts, which were fragmented, are consolidated which resulted in a 20% gain in overall productivity. Centralizing translation memories, terminologies as well as standardizing translation

guidelines and processes also provides the necessary asset management foundation for training MT system. The next phase of the strategy is then, to replace most of human translation with MT post-editing (MT-PE) to the extent possible. In the future, once the engine has reached a certain level of performance through incremental customizations, the plan is to utilize the same MT system for 'best-effort' translation to provide users with the option to potentially view any TechDoc with MT 'gisting' quality, further uplifting the user experience.

2. MT Post-Editing Solution

Cisco is currently implementing the MT Post-Editing scheme as part of the second phase of the overall MT strategy as outlined in the previous section. Following describes the key components of the end-to-end solution that was deployed to enable MT Post-Editing:

- **Workflow Infrastructure:**

  The latest hybrid MT system from SYSTRAN Software, hosted on the cloud, was integrated with SDL's translation management system (see Figure 2.1). Model training and asset preparations are mostly performed through SYSTRAN's self-service tools (corpus, dictionary and training manager). Additional scripts have also been developed to handle data cleanup before training, and to replace frequent errors at the MT output.

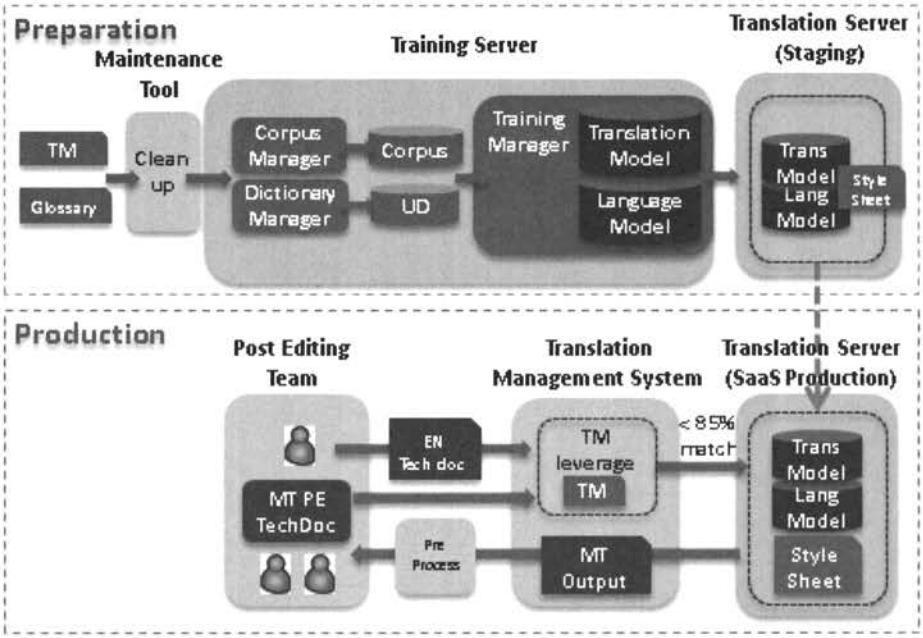


Figure 2.1: Training and Production Workflow for MT-PE

- **Customization Process:**

To continuously adapt the engine, we created a monthly maintenance process, where assets from completed translations were recycled and feedback from post-editors reflected into translation memory and user dictionaries. User dictionaries were also updated and fixed with the output of auto-generated dictionaries,

semi auto-generated DNT dictionaries and Cisco glossaries. Model training was then executed, evaluated to see whether the quality improved and could be considered to replace the previous model. Some spot customization at engine level were also performed, like improvement of style sheets and adding rules for data trimming during training.

- **Translation & Post-Editing Process:**

English source documents are submitted to TMS where appropriate workflow and TM matches are applied. Segments with lower than 85% matches are sent to the customized MT system. Finally the combined MT output and TM matches are supplied as bilingual files to translators for post-editing.

- **Post-Editor Training and Feedback:**

An important element of success is to provide necessary training to the post-editors to fully integrate them with the MT-PE operation. The training includes briefing on the background and overall characteristics of MT technology, and explanation of a post-editing guideline showing common errors, and how to avoid over post-editing etc. Feedback from post-editors is collected upon completion of a MT-PE project , such that the MT team, can analyze the comments and identify potential areas of improvement.

3. Performance Results

Figure 3.1 shows the evolution of the GTM & BLEU scores of the English to Japanese MT system, as a result of the incremental customization effort.

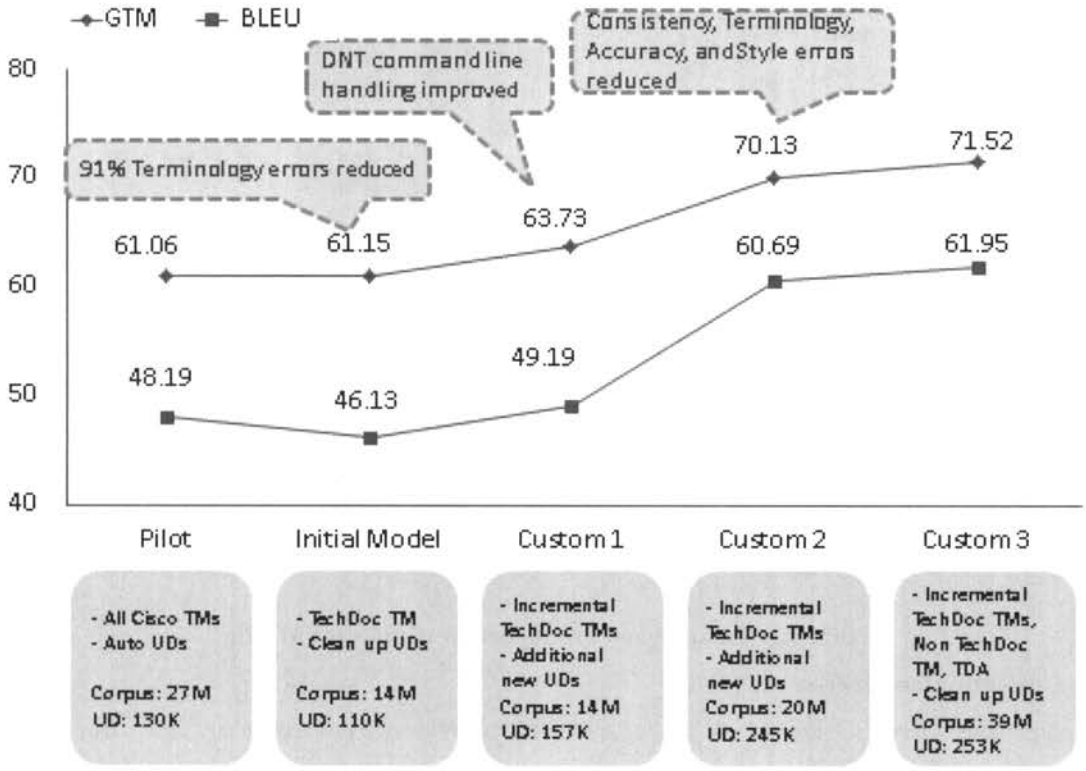


Figure 3.1: Automatic Evaluation Scores

The highest score in this automatic evaluation was obtained (61.95 for BLEU, 71.52 for GTM) after four cycles of customization using 39 million words for hybrid training and a user dictionary of 253K entries. The use of clean TMs and continuous dictionaries updates were effective in reducing terminology errors by 91% from the initial model and increasing the accuracy. In model "Custom 3", TMs from different content type (non-Techdocs domain), and third party data (selected TMs from the TAUS Data Association) were included in the training– surprisingly the score improved by a little (we initially thought that adding non-Cisco data and content from another content type would degrade the performance but that was not the case). In general, it was observed that the statistical component of the hybrid engine significantly improved the fluency and style. This is critical for the readability, considering that translated documents are made up of mixture of TM matches from past human translation and post-edited segments.

In terms of MT-PE operation, so far, 40% of all translations are going through this process. Overall productivity improvement for post-editors has ramped up to 45% as the quality of the engine increased over time and post-editors became more used to the process.

Initial feedback from the post-editors is positive – the impression was that the quality was higher than expected for short sentence (<15 words). We forecast conservatively 34% of translation savings, after the system is upgraded in near future, to handle remaining integration issues between MT and TMS for 'highly tagged' contents, such that post-editing can be applied to those documents without loss of productivity (ex: proper DNT interpretations for command/GUI, tag positioning).

Some early evaluations were conducted with model 'Custom 2' to assess whether the level of 'gisting' quality of the trained system is 'good enough' to be used as-is for real-time translation. Evaluators were asked to rate the raw MT output on randomly selected sets of phrases (300 words) within an 'understandability' scale of 1(lowest) to 5 (highest). For this test, the average quality score was 4.35, but it dropped to 3.72 for test set containing longer phrases. Nevertheless, the overall results are still encouraging, and quality perception is generally higher compared to the two public engines (Google, Bing). More extensive evaluations on a broader test set are needed to be able to confirm the consistency of the quality level before concluding on its usability for on-demand translation purpose.

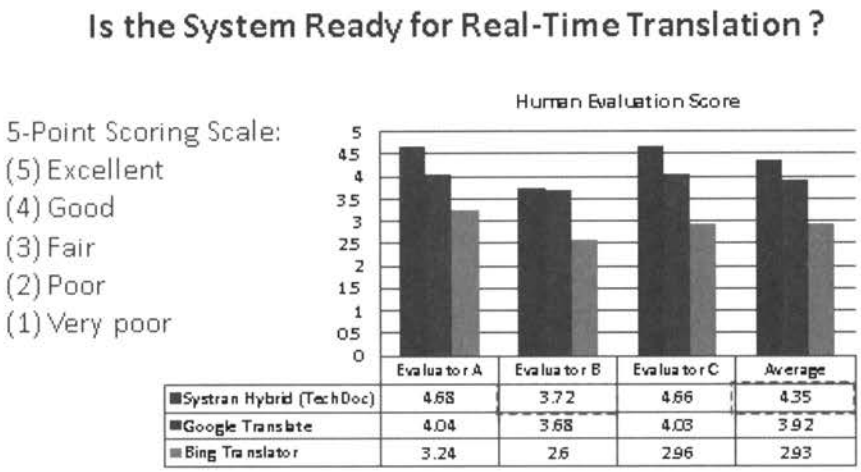


Figure 3.2: Comparison of gisting quality

Finally it was noted that directly adding more TMs for training does not necessarily increase the quality. In fact, when including corpus from Cisco support material, the score dropped from 4.35 to 3.73. Similarly, the best score is obtained for knowledge base (Tech Support) content when the system was trained with 'in-domain' TMs.

MT System		TechDoc	Tech Support
Customized (Training Corpus)	Sytran Hybrid (20M TechDoc)	4.35	2.78
	Systran Hybrid (4M TechSupport)	3.81	3.11
	Systran Hybrid (24M TechDoc+TechSupport)	3.73	2.85
Public	Google Translate	3.92	2.99
	Bing Translator	2.93	2.33

4. Future work

As part of future effort, we will look to integrate MT with the content delivery system infrastructure in support of on-demand translation. We will continue to try out different approaches to further fine-tune the MT system - testing with more combinations of options settings and adding additional corpuses in the training. Finally, as we look to scale the MT solution to handle various types of contents and to optimize training – we will explore the

concept of 'dynamic adaptation'. The idea is to cluster all the TMs and based on the input content , the system will select the most appropriate set of TMs (that is highly correlated with the content to be translated) that will be used for training. The expectation is that this data-based approach, if successful, could deliver the 'best' translation quality with the available training data, potentially reducing the amount of required human effort for manual evaluations of training combinations.

Project Report

Collaborative Machine Translation: You, Your community and Microsoft's knowledge working together

Chris Wendt
Microsoft Research – Machine Translation
christw@microsoft.com

Introduction

The language quality of automatic translation today is at a level we can consider sufficient, or "good enough" for many use cases and language pairs. In addition to using automatic translation in a post-editing workflow, the scenarios for unedited use of machine translation output include real-time chat, technical articles written in controlled language, or the consumption of foreign language web pages, documents, social media and travel situations. A generally available MT system can do quite well in these scenarios, applying the most common used phrases and expressions, on any run of text for which it is able to parse and match to its data and algorithms.

We have some options to improve on the situation: The quality of an MT system can be increased drastically by customizing it to the terminology and style of the subject of the text to be translated, be it a general area like agriculture, software technology, automotive etc., or, and maybe in addition, the terms and phrases that an organization prefers to use in translation, for instance product names or the things that these products do. An area of style and terminology is often called a *domain*. By reducing the ambiguity inherent to language, and targeting the system to a certain specialization, we can see a significant improvement of any quality score over uncustomized systems.

MT systems that are using statistical methods at its core, train their probabilistic models on previously translated documents, and possibly additional documents in the target language. Microsoft Translator is a statistical MT system. In the following we discuss the methods of training a customized MT system that Microsoft Translator exposes.

SMT Quality Factors

The relevant factors in the quality of a statistical machine translation system are algorithms and data.

Statistical MT = Algorithms + Data

Algorithms
- Decoder: the decoder is the core of the translation engine, and is the piece of code that performs the actual mapping of a source sentence to a sentence in the target language. The features of Microsoft Translator that are especially relevant for customization to a domain are:
  - Multiple translation tables
  - Multiple target language models
  - An efficient training system
  - A scalable, reliable and fast runtime
- Pre- and post-processing logic that allows the decoder to consider linguistic

attributes of the source text.

Data

- Parallel documents: A body of documents, each document in two languages, is the most important training data for customization. It teaches the system how to translate phrases.
- Target language documents: a body of documents in the target language, teaching the system how things are commonly said in the target language – it helps pick the right context and inflection for a word that is being considered as a candidate.
- Domain-specific parallel documents: documents that teach the system the preferred translation for a specific area of terminology and style, for instance agriculture, and the preferred translations for the organization that performs the customization.
- Domain-specific target language documents: documents providing examples for the appropriate use of the domain-specific phrases. The system learns the right context and inflections within the given domain from these documents.
- User edits, votes, ranking: community feedback on the translations that the system produced. This includes votes and ranking by the audience for the unedited machine translation, as well as corrections by professionals and amateurs. This may include human translators doing selective or complete post-editing, visitors of the site, employees, partners, fans, enthusiasts, family, contractors – anybody who is motivated to give feedback. The key is that this feedback is tied to the domain in question.
- Tuning set: a set of sentences in original and translation, used by the system to adjust its weights and parameters during training. The tuning set and the test set should be carefully selected to optimally represent the domain-specific material to be translated in the future.
- Test set: a set of sentences in original and translation, which generates an automatic quality score, and is ready for human inspection to judge the quality. Neither the test set not the tuning set should have any overlap with each other or with the data used for training, in the actual sentences used.

In the following, we will further inspect the effect of multiple translation tables, multiple target language models, and the methods by which the community can influence the behavior and quality of the customized MT system.

Collaborate to Customize

Microsoft Translator features a built-in translation memory (TM), that is fed by the simple AddTranslation() method in the API, and acts on sentences. At time of submission, the application decides on the rating of the given sentence pair based on the authority of the person providing the pair. The system accepts votes as well as edits, where a vote is simply an increment of a counter for the given pair, using the authority of the voter to determine the rating. An application may assign an authority between 1 and 10 to its user, where the MT system produced results always have a rating of 5.

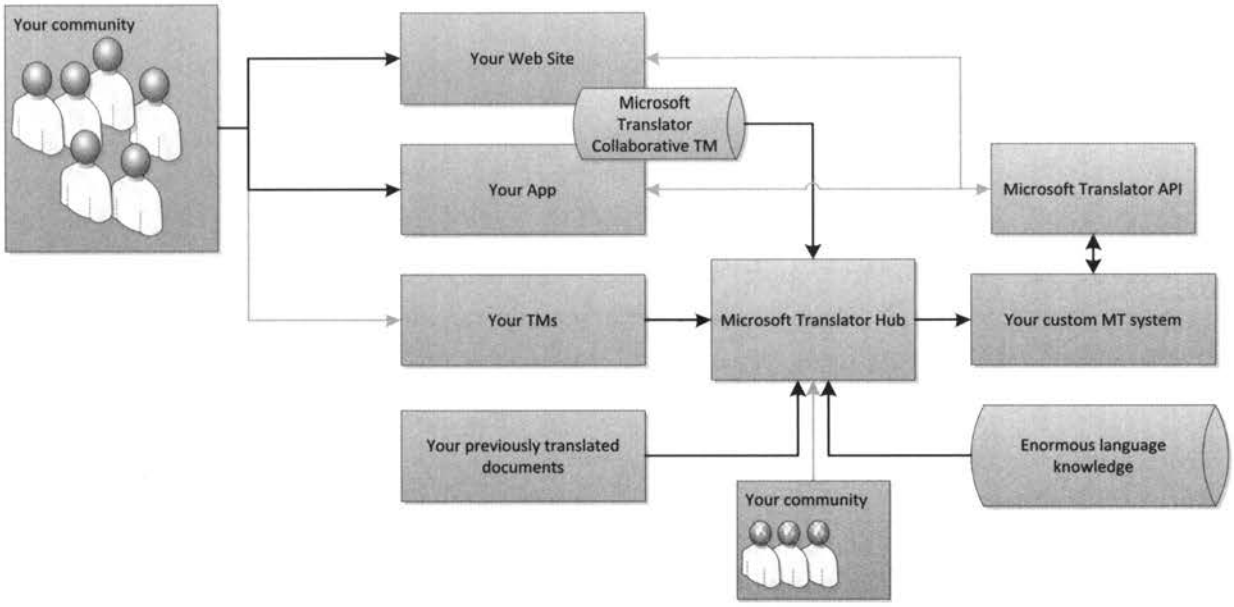
During translation, Microsoft Translator returns the highest rated translation. This will be the highest rated 100% matching entry in the translation memory, or it will be the result of automatic translation, if there is no entry with a rating higher than 5. The Microsoft Translator API also provides a method

to retrieve all entries for a given source sentence, to allow the application to show a voting and ranking UI to the user.

This built-in translation memory serves two purposes: 1) It collects edits, corrections and votes from users for the purpose of using this data in training, and b) it has an immediate effect on the appearance of the document on the site, and all subsequent translations of the same sentence, hereby providing instant gratification to the community: instantly improved translations for the edited documents as well as any boilerplate text with frequent occurrence. The improvements exposed by the translation memory are immediate, whereas the customized system training is neither certain nor immediate, but has an effect on the quality of many more sentences.

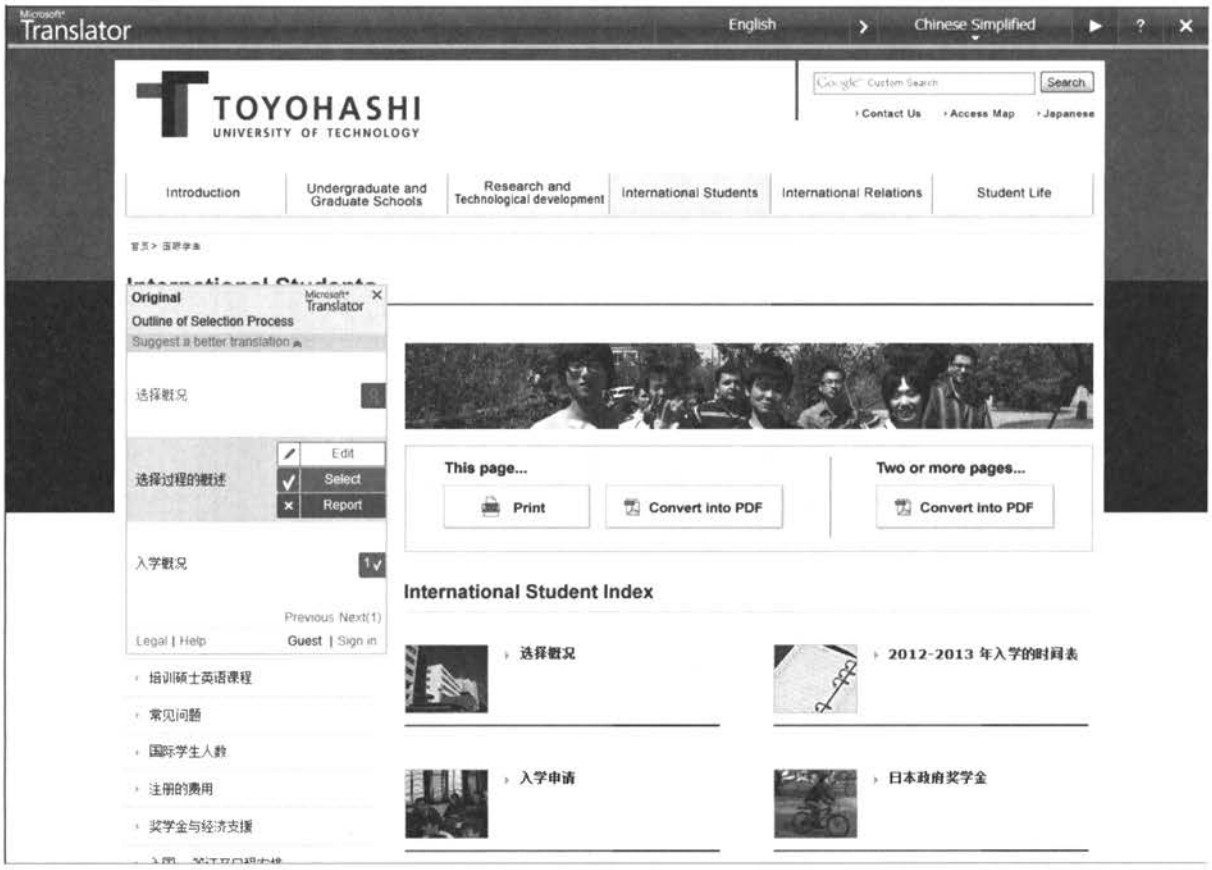
Microsoft Translator Hub is a subsystem within Microsoft Translator, which features functionality to

- invite and manage collaborators in the customization process,
- upload exiting TMs in TMX format, previously translated documents in a variety of formats including PDF, Microsoft Office and plain text,
- Invite and manage reviewers or translators for uploaded and translated documents,
- import data from the collaboratively created TM,
- select the appropriate tuning and test sets from the uploaded documents,
- initiate training, inspect the results, and eventually
- deploy the customized system, making it available via the Microsoft Translator API.



Example: Collecting Feedback on a Web Site

Owners of a web site, instead of writing their own application to collect edits, corrections or votes form the visitors, may choose to deploy the Microsoft Translator widget , which does all of this for them, and provides simple user management and bulk editing features. Here is an example of how this can look like:



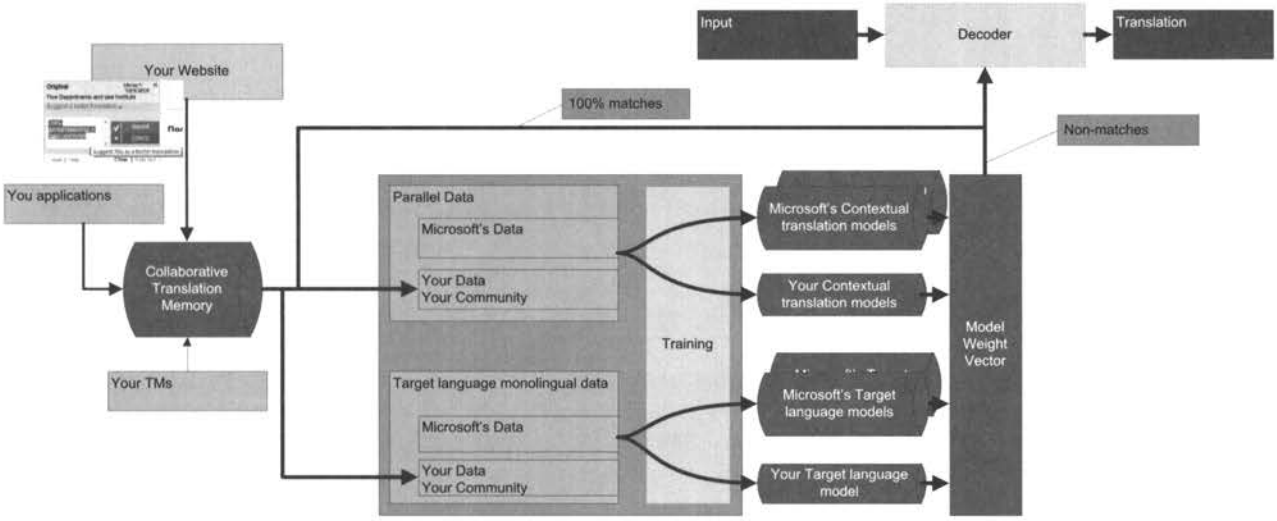
The illustration shows the web site of Toyohashi University of Technology, in fact their original English language site, with an automatic translation to Chinese. Visitors to the site can hover the mouse over any portion of the translated text, see the original text, and have the option to vote for or edit any of the translation alternatives. And of course have the option to identify themselves, so that they can do all of this under their own authority. If they have sufficient authority, any voted translation and any edits become the translation that any visitor of this site sees from now on, as the master translation.

A Look Inside: The Tech Behind the Microsoft Translator Hub

When a project owner in the Microsoft Translator Hub initiates a training, the following things happen:

- The Hub builds a contextual translation model from the parallel data that the owner provided.
- The Hub builds a target language model from the provided target language material.
- The Hub uses the tuning set to generate the optimal model weight vector, which assigns the weights to the each of the custom models as well as each of the models that Microsoft Translator provides generically, including Microsoft Translator's own contextual translation and target language models.
- The Hub translates the test set with the optimal weight vector, and generates a score.

After deployment, all of the models and the weight vector become part of the Microsoft Translator API, and are accessed via a private category key, which is provided to user by the Hub.



At runtime, any translations (input) that the system performs, are first matched against the collaborative translation memory, and in the case of no match with an "approved" rating, are translated by the customized MT system.

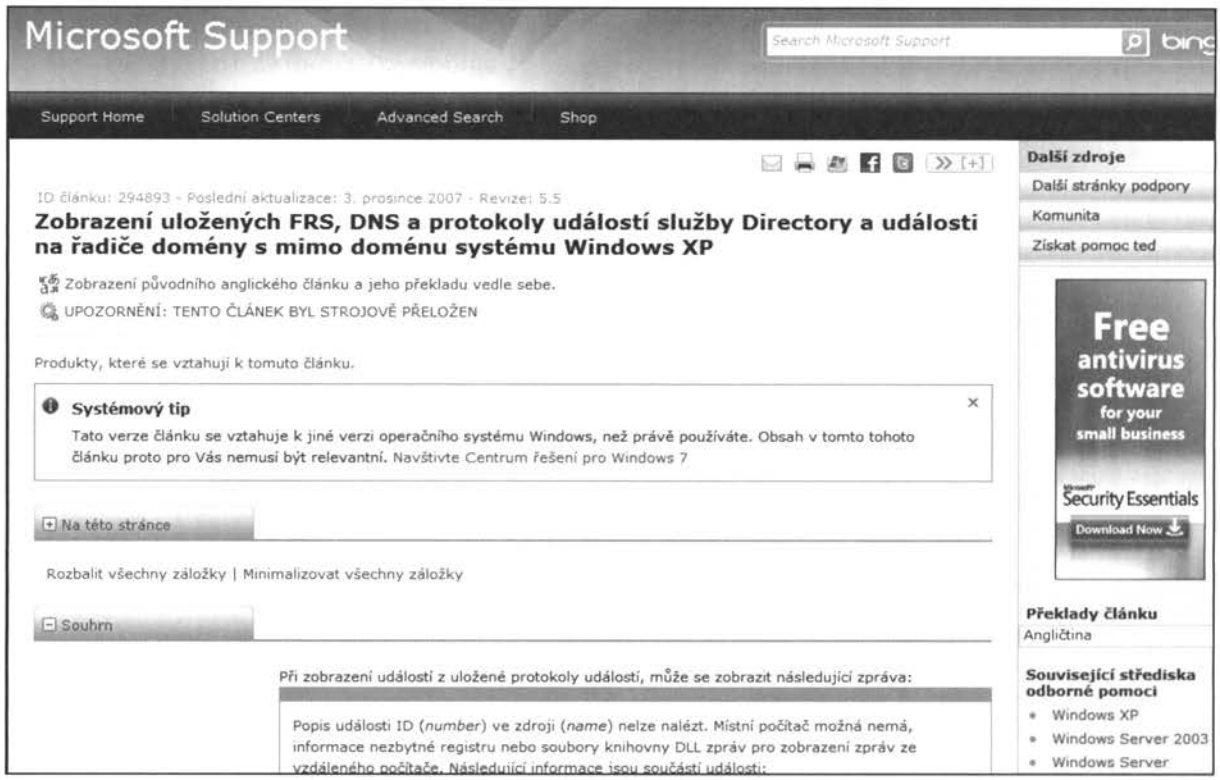
The diagram shows that the community data serves the purpose of immediately serving the corrections and approvals for exact sentence matches, and is also available for repeatedly training the custom MT engine, in self-service fashion, adding new domain-specific training data in each iteration.

Customization In Practice

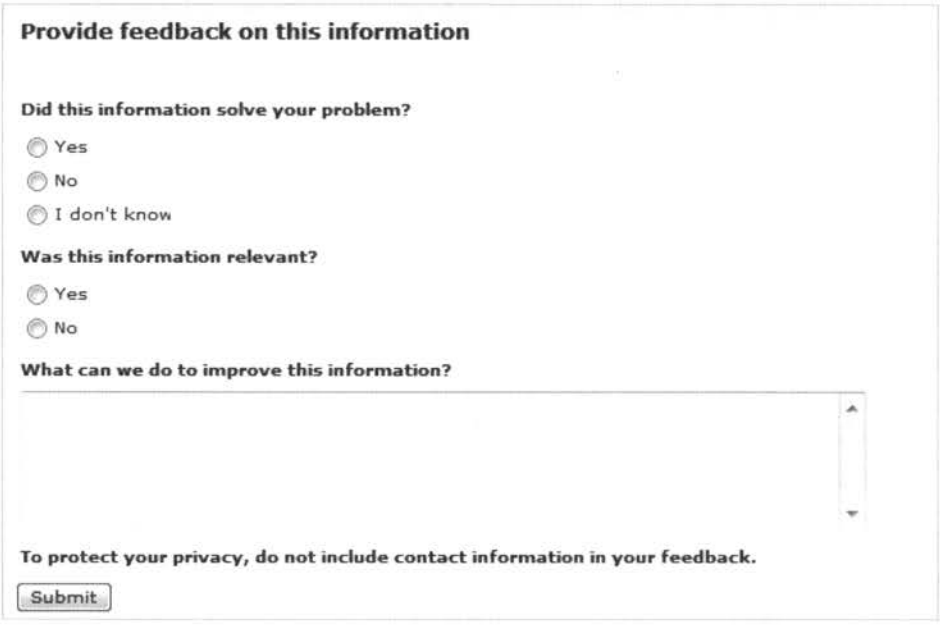
Microsoft itself has been using a customized MT system with the knowledge base since 2003. Support agents around the world have been entering corrections via a viewer with editing functionality for many years, and now the functionality is also available to the general public.

Microsoft translates approximately 10% of the knowledge base content by humans, the remaining 90 percent are published as unedited machine translation. Which articles get translated by humans is determined by the number of page views received on the machine translated article. The presence of both human translated and machine translated articles within the same corpus and the same languages, gives us a good metric for comparing the effect of the published articles on the customer seeking to solve a software problem.

This is how an automatically translated article is presented to the visitor. Note the clear indication that this is an automatically translated article, near the top.

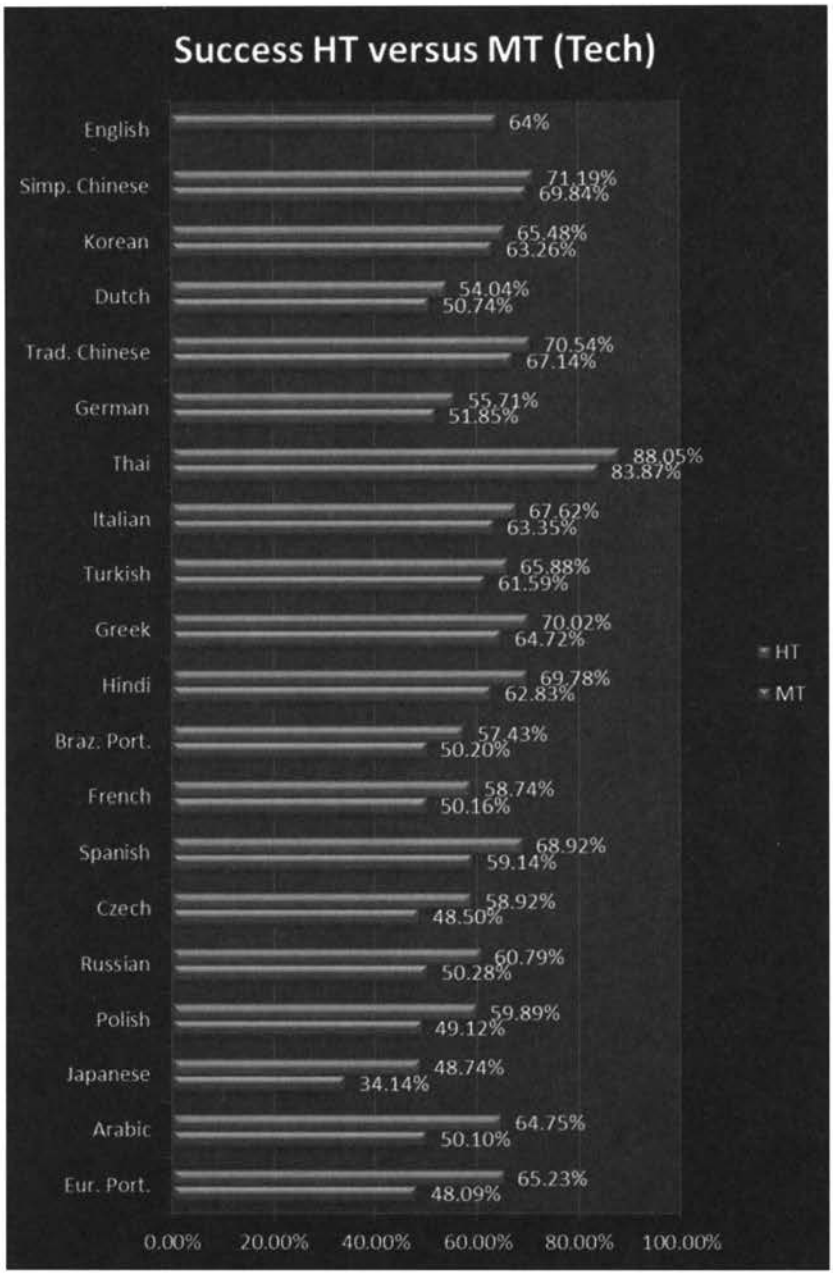


At the end of every article, regardless of language or type of translation, user is asked the following questions – in their own language:



The success rate is the ratio of people who answer "Yes" to the question "Did this information solve your problem". Note that the site is not asking if the visitor liked the translation or if it was accurate, only if it solved the problem. This is important: the applicability of automatic translation needs to be tested within the context of the scenario to be addressed. It is extremely difficult to map a quality score gained outside the scenario to the applicability within.

The following table, provided by Martine Smets in Microsoft's customer support organization, shows the success rate of the articles that have been translated by a human (HT-red) vs. the articles translated by automatic translation (MT-blue).



In Conclusion

Data is the Key to MT Quality: Collaboration makes it real.

Your TMs and your community feedback are excellent sources of in-domain training data. You can achieve good quality by combining it with Microsoft's big data – statistically weighted. The Microsoft Translator Widget collects data from your community. Microsoft Translator Hub collects data from you and your community, to train the best system possible. You may use the collaborative translation TM for instant effect, and instant gratification.

References

Microsoft Translator: http://www.microsoft.com/translator

Microsoft Translator Hub: http://hub.microsofttranslator.com

Microsoft Translator API: http://api.microsofttranslator.com, with links to forums and support

Project Report

The Power of MT

---What Yamagata MT does for Honda Motor Europe---

Yamagata Europe

Heidi van Hiel

概要

YAMAGATA ヨーロッパは、ホンダ(Honda Motor Europe 社)からの依頼を受け、一連の業務革新の一環として 2 種類の機械翻訳(MT)プロジェクトに取り組んでいる。このレポートでは、2 つの MT プロジェクトを遂行するうえでの問題点とその対応策、プロセス、そしてその結果を説明する。最初のプロジェクトの課題は、ヨーロッパのディーラーがそれぞれの言語で書いたものを 12 時間以内に他言語から英語に翻訳することであり、2 つめのプロジェクトの課題は、多言語での同時チャットサービスである(前編集/後編集なしで 90%理解できるものに！)。いずれも(ディーラーを含めた)ホンダ社内データベースに組み込まれたものであり、社内での使用に限られたものである。(中村)

This article describes two MT projects that Yamagata Europe has realised with Honda Motor Europe. For each project, we describe the specific hurdles that had to be taken and the results that were achieved. The first project is a project where data is translated into English within a 12 hour scale, the second project is a sample of an instant multilingual chatting service. Both systems are for in-house use at Honda Motor Europe only and fully integrated in the Honda IT systems.

Project 1: Honda warranty claim project

Over the years, Honda has built up a database with valuable information about all kinds of warranty-related quality issues: whenever a Honda dealer (car, bike, power equipment) is confronted with a problem about a device part that is still in warranty, the dealer is requested to fill in all information about this warranty issue into Honda's warranty database (problem description, problem diagnosis, reparation details, etc.). The dealers can input this information as free text in their own language.

In July 2009, we were asked by Honda to create a 'translation system' to generate English translations for text from this warranty claim database. The request for translation came from the Honda Product Improvement center in the UK: they wanted to make more and better use of the information in the warranty database.

Three languages were selected to be translated into English with MT: German, Italian and Russian, because these languages represent the three major European markets (over 50% of all warranty issues). In the future more languages might follow.

Source text quality

A closer look into the contents of this database taught us that this project would be very challenging: the quality of the source text was the first hurdle to take. Since every dealer was allowed to input the data in his own way (including grammar mistakes, spelling mistakes, stylistic differences, punctuation issues, abbreviations, typical Honda lingo), the source text as such could not be used as input for a MT engine.

Poodle and the importance of pre-editing

To handle the data, we designed a translation system called "Poodle": a combination of semi-automated pre-editing, TM-integrated machine translation and ultra-light post-editing.

The first step of the Poodle translation flow is to run the text strings through the translation memory. Since we have been building up a huge translation memory over the past years, on average 30% the claims can be translated using the translation memory. The remaining 70% will have to be translated via MT.

As explained before, the quality of the remaining text needs to be improved before being sent to the MT engines. The [Finetune source] function of Poodle takes care of this. The following actions are executed:

- Replacement of (unusual or uncommon) abbreviations by their fully written counterparts
- Various replacements (including replacement of incorrectly used punctuation) via regular expression replacements
- Formatting corrections (correction of casing (uppercase – lowercase), punctuation, spacing, bracketing, usage of hard/soft returns etc.)
- Spellcheck

Most finetune actions are executed completely automatically, but some require human intervention, such as the spellcheck. Here are some examples to show the difference between the original source and the pre-edited source:

Source before finetuning (German)	**Source after finetuning**
KEIN BLUETOOTH MOEGLICH, SN 242001955	Kein Bluetooth möglich, SN 242001955.
PRÜF. AM FHZG.	Prüfung am Fahrzeug.
TIEFER EINSCHNITT IN REIFEN HI AN NEUFZG	Tiefer Einschnitt in Reifen hinten an Neufahrzeug.
FENSTERHEBER A+E U ERN.	Fensterheber Ausbau und Einbau und erneuern.

Generating MT output: Systran MT and Poodle QA

Once the source text is finetuned, we can move to the most important step: generating an MT output. Yamagata Europe has generated Warranty MT engines, using Systran technology. The finetuned source strings are sent to the MT engines, and a translation output is generated. The following screenshot shows some MT results. The left column shows the finetuned source, the right column shows the MT output.

Finetuned Source	Translation
Beschichtung der Windschutzscheibe löst sich	Coating of the windshield releases itself
Kein Bluetooth möglich, SN 242001955	No Bluetooth possible, SN 242001955
Sichtprüfung am Fahrzeug mit der Kundin	Visual inspection at the vehicle with the customer
Windgeräusche aus dem Türspiegelbereich rechts	Wind noises from the door mirror range on the right
Knarzgeräusch beim einlenken an der Vorderachse	Creaking noise when cornering at the front axle
Tiefer Einschnitt in Reifen hinten an Neufahrzeug.	Deep cut in tires rear at new vehicle.
Fensterheber Ausbau und Einbau und erneuern	Regulators dismounting and installation and replacement
Rußpartikelfilter ist verstopft	Particulate filter is clogged
Bremsen hinten machen Geräusche	Brakes rear make noises

The engines were heavily trained to cope with the warranty source text: huge glossaries were extracted from existing translations memories, parts lists and warranty legacy data.

After translation, a light and automatic post-editing is applied. To locate possible mistranslations in the output, a short QA check is applied. The following checks are executed:

- Partially forgotten translations: check for untranslated words. In case a word is unknown to the MT engine, it is tagged as untranslated. The tagged strings are automatically captured at Honda Motor Europe and translated internally. The translated strings are delivered back to Yamagata Europe, and added to the glossaries of the MT engines. This ensures that the these strings will always be translated in future files.
- Negations check: a typical mistake of a hybrid MT engine is to translate a negative sentence as an affirmative sentence or vice versa. This check looks for negation elements in source and target and reports a warning in case of a mismatch between source and target.

- Number check: this check compares numbers in source and target and reports a warning in case of a number mismatch between source and target.
- Check on empty translations: this check searches for sentences that were skipped during the MT processing.

Errors spotted during QA check have to be corrected manually. When all errors are solved, the final translated file is generated, and the Translation Memory is updated with the MT translations.

Measuring the quality

The warranty project went 'live' in December 2010. During the technical developments, we also worked out a methodology to measure the quality of the MT-output based on understandability levels. This enabled us to report progress to the Honda higher management. Since the warranty data is used only internally within Honda (no print or publishing), the goal of this project is to generate "understandable translations". This means that the translation does not need to be grammatically or linguistically perfect, but it needs to convey the meaning of the source text correctly.

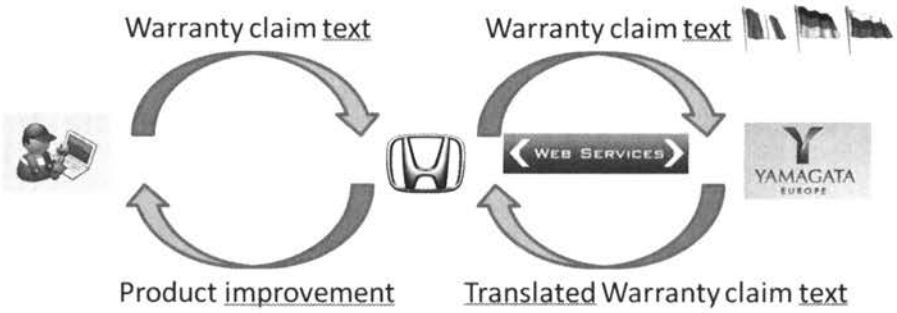
Although we use BLEU (Bilingual Evaluation Understudy) scores during engine training and evaluation, we developed our own quality levels for the evaluation of this project. A low BLEU score does not necessarily indicate low understandability, but it mainly indicates the effort that a post-editor will have to make to convert the translation into a perfect 'human' translation. However, for this project, we did not want to evaluate the translations on their perfectness, but on their understandability. Hence, to score the understandability of the MT output, we developed 4 quality levels:
1: Not understandable (Clearly no good and doesn't make sense).
2: Barely understandable (Most words translated, but general meaning not clear or needs really thoughtful interpretation).
3: Fully understandable (Can understand meaning, but grammar is not correct, may need a little interpretation).
4: Perfectly understandable (No interpretation required or grammatically correct text).
Based on these 4 quality levels, we reached an understandability of 87% (levels 3 and 4), which is 2% higher than the target set by Honda (85%).

The translation highway

Next to the development of pre- and post-editing tools and the training of the MT engines, we worked on IT integration with the Honda team, to ensure smooth and automatic file transfer via web services: the 'translation highway'. A web service is a method to enable communication between 2 electronic devices over the internet: a web service uses XML-based messages that are sent over Internet protocols to support direct interaction with other software applications. Web services have many advantages, such as platform/technology independency, low communication cost (usually SOAP over HTTP protocol), support for other communication means (e.g. web service over FTP), etc.

For the Warranty project, Honda makes 2 exports (xml) from their warranty claim database per day. The exports are automatically transferred (web services) to a server at Yamagata Europe. The translated files are sent back on the same day, via the same way. This means: no email traffic, no telephone communication, but a completely automatic data transfer.

The following picture describes the flow:



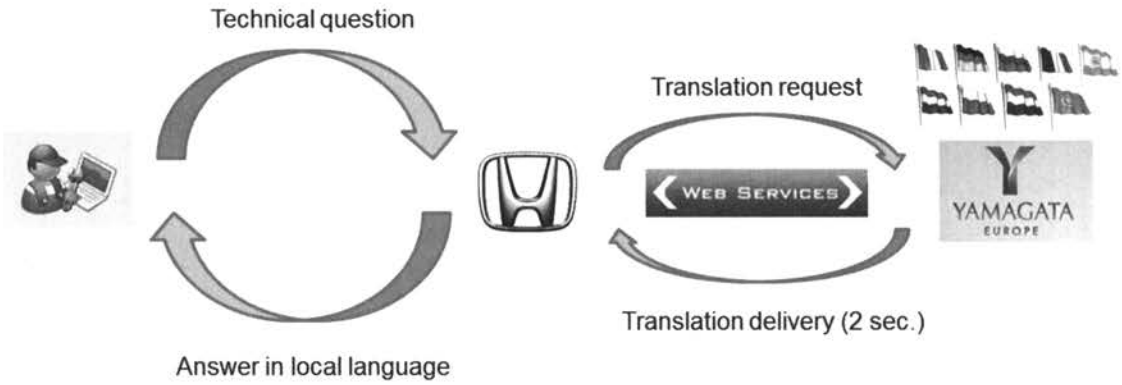
Project 2: Honda online chat project

Even before going 'live' with the first project, Honda was enthusiastic about the MT potential and organised an internal brainstorming to see if other projects could link to this new 'translation highway'. The result of this brainstorm fitted well into a planned restructuring of the Honda Motor Europe service support to the local dealers. This support service used to exist in each European country, but was not affordable anymore giving the growing number of new markets (countries) and engine technologies. Honda was already planning a web-based communication and chat interface between the centralised support service center in Offenbach and the in-country dealers. Honda also decided to add a "translation request button" to the interface in order to allow users to request a translation into English or into their native language.

Full speed on the translation highway

The main difference between the online chat project and the warranty project is speed: while the warranty project requires same day delivery of the translations, the online chat project requires immediately delivery (within 2 seconds). This means there is no time for human intervention at all: no pre-processing, no post-processing. On top of this, Honda set an ambitious target and requested a 90% understandability.

The following picture describes the flow:



The currently supported languages (always in pair with English) are: Dutch, French, German, Hungarian, Italian, Polish, Portuguese, Russian, Spanish. The project went live in February 2012.

Pre-editing training: preparatory human intervention

Since the chat project does not allow any human intervention in the MT process, we decided to train the engineers in the Honda support service center about how to optimise the source text in order to get a nice and understandable MT output. An intensive 2-day training was given in order to show the importance of a good source text. Simplified and structured source writing instructions can make a huge difference in the MT output, some examples will clarify this:

➔ Instruction: Write short sentences:

Pre-editing action	Source text	MT output	Quality rating
-	Lackteil sonst nicht beschädigt, Teil muss erneuert werden.	Paint part otherwise damaged, part does not have to be replaced.	2. Barely understandable
After pre-editing (sentence split in 2 separate sentences)	Lackteil ist sonst nicht beschädigt. Teil muss erneuert werden.	Paint part is not damaged otherwise. Part must be replaced.	4. Perfectly understandable

➔ Use correct casing

Pre-editing action	Source text	MT output	Quality rating
-	Fremder laut im Kofferraum	More strangely loud in the trunk	3. Understandable
After pre-editing (Laut -> Laut (German nouns have to be capitalised))	Fremder Laut im Kofferraum	Strange sound in the trunk	4. Perfectly understandable

➔ Limit the use of the gerund (English –ing form)

Pre-editing action	Source text	MT output	Quality rating
-	Vehicle battery keeps going flat.	Fahrzeugbatterie hält, flach zu gehen.	2. Barely understandable
After pre-editing (keeps going flat -> is constantly empty)	The vehicle battery is constantly empty.	Fahrzeugbatterie ist ständig leer.	4. Perfectly understandable

MT backup: Human translation button

The Honda requirements on this project are very high: translation delivery within 2 seconds with 90% understandability. This is because the Honda dealers used to have immediate telephone support in their own language for most of the issues they encountered.

At this moment, we have reached an average understandability level of 79,5% - so we still need to do better. But in the meantime, we have integrated a human translation system to the MT flow: if the MT output is not satisfying, the engineer in the Technical Support center can press a button to request a human translation. The request is sent to Yamagata over the translation highway and the human translation will be delivered back to Honda.

Conclusion: Pre-editing is the key element for good MT output

The biggest challenge for both Honda projects was the unstructured source content. With a large user group as in the Honda case, it is almost impossible to control and manage the data and text input. On the other hand, we

have experienced that controlling the source is crucial to obtain understandable MT output, this is why we have put the focus on pre-editing rather than on post-editing. This means that training the MT engines was only a small part of this quite complex project: controlling or upgrading the input was the biggest challenge.

シンポジウム参加報告

TAUS Executive Forum Tokyo 2012 報告

立見　みどり

2012 年 4 月 19 日、20 日の 2 日間にわたり、日本オラクル本社のあるオラクル青山センターで開催された「TAUS Executive Forum Tokyo 2012」に参加しました。副題は「Translation in the 21st Century」というもので、より多くの言語で、より動的に、そしてユーザーを中心に据えた自動翻訳の提供と利用をテーマにしたフォーラムでした。参加者は事前登録数で 60 名を超える盛況で、海外からの出席者も多く、20 以上あった発表はすべて英語で行われました。

TAUS は、Translation Automaton User Society という名のとおり、MT（自動翻訳／機械翻訳）システムの実際の利用に焦点を当てた団体であるため、フォーラムの出席者も、自動翻訳を利用する側の企業（Buyers of translation）と自動翻訳システムおよびその周辺の技術とサービスを提供する企業（Providers of translation）の方が多数を占めます。このため、学術的なコンファレンスとはやや雰囲気が異なり、まさに現場からの報告という感じです。質疑応答も、各参加者が日々の業務で直面している問題を中心とした内容で、非常に活気ある 2 日間でした。

初日のオープニング・プレゼンテーションでは、TAUS のディレクターである Jaap van der Meer が翻訳を取り巻く今後 5 年間の変化として、翻訳メモリが個人的または特定の企業専用のものから、より公共の共有財産へと整備されることになるだろう、また自動翻訳やクラウドリソースの活用により、言語のプロフェッショナルである翻訳者の仕事はより高い品質が求められる業務に絞り込まれていくだろうと予測しました。またこれに関連して、翻訳に求められる品質も画一的なものでなく、必要性や提供可能性に応じて品質を選択することが当たり前になっていき、翻訳にかかわるプロフェッショナルとノンプロフェッショナルの幸せな住み分けを実現することが課題となると語りました。

続いて、AAMT 会長の井佐原均氏がアジアの学術界と商業界における機械翻訳関連の活動状況、および人手による翻訳を助けるための用語抽出ツールや翻訳候補を列挙するツールの開発努力について紹介しました。その後、国立国会図書館前館長の長尾真氏が図書館資料のデジタル化についての現状について話し、中国や韓国の国立図書館との協力により、目録情報や資料のコンテンツそのものを中国語や韓国語に機械翻訳する試みについても語りました。

引き続き多数の発表がありましたが、その中からいくつかを以下に簡単に紹介させていただきます。

ユーザー企業の視点から

ユーザー企業の視点からは、現在利用可能な MT システムを自社のニーズに合わせて効果的に利用するための試みやシステム作り、評価などについての発表がありました。PayPal の Catherine Dove 氏は、顧客向けコンテンツを、品質を維持しながら費用効果の高い方法でローカライズするため、① Acrolinx（オーサリング支援ソフトウェア）を使って英語の原文の品質を向上させる。② MT の性能自体を向上させる。③ ポストエディティング（後編集）のルールを導入する。などを行い、生産性が向上したと報告しました。

日本オラクルの俣野宏子氏と斎藤玲子氏の発表では、まずMTシステム利用の方法として、① MT出力をそのまま使用する、② 社内の人材によるライトエディット（正しい訳文にする）を行う、③ 社外の編集者によるライトエディットを行う、④ 社外の編集者によるフルエディット（正しく読みやすい訳文にする）を行うという4つの選択肢が提示されました。その後それぞれの選択肢をスピード、品質、コストの面ら評価した上で、どの方法がどのような場面（無償で提供する大量の文書、分量の少ないテキスト、オンラインで提供する定型文の多い大量の文書、高品質の翻訳が求められる文書など）に適しているかが検討されました。また、ポストエディットを効率よく行うための心構え、手法、品質ガイドラインなども提案されました。

Autodeskの Mirko Plitt氏は、同社ですでにいくつかの言語で導入されている SMT＋ポストエディティングというワークフローを日本で導入した場合の生産性テストについて発表しました。それによると、2010年のテストでは MT を利用することで生産性が落ちたため導入を断念。しかし日本語の語順に合わせて英語の原文の語順を自動的に並び替えるプログラムを導入したところ、2011年のテストでは大幅な生産性向上が見られたとのことでした。

プロバイダ企業の視点から

翻訳メモリ、MTシステムに加え翻訳サービスも提供している総合的な翻訳プロバイダである SDLの Daniel Marcu氏は、SMTの品質を上げる要素として、優れたアルゴリズムや豊富なトレーニングデータ以上に、利用側からのフィードバックデータやドメイン固有のデータが重要であると語りました。

翻訳ベンダーの大手である翻訳センターの河野弘毅氏からは、独自の CAT ツールを利用、カスタマイズ、維持していく上でのさまざまな問題点（ファイル形式の互換性やドメインごとに作業手順が異なることなど）や、課題（SMT 用トレーニングデータの集積、クラウドに対するクライアント側の理解を得ること）などについて、同社の経験と苦労に基づく話がありました。

オーサリングソフトの開発会社である Acrolinx の柳英夫氏は、日本語の原文を特定のルールに基づいて執筆することで英語への MT の品質を上げる試みについて発表しました。この試みは、プロのテクニカルライターではない、エンジニアが技術文書を執筆する場合を想定したもので、明確に定義されたルールに従って原文を記述することで、MT品質の向上に一定の効果がもたらされる可能性を示しました。

また、SDLの Alan Chung氏と Ciscoの Dieu Tran氏は、プロバイダとユーザーが共同で MT＋ポストエディティングソリューションを開発し、コスト削減と納期の短縮を実現した事例を報告しました。

クラウドとコラボレーション

従来のプロバイダ企業とユーザー企業の取引という形を超えて、個人や組織が小さな単位で、あるいは必ずしも金銭のやり取りが発生しない環境で、サービスやリソースを提供したり利用したりする枠組みについても発表がありました。

バオバブの相良美織氏は、同社が提供する、MTシステムと留学生というリソースを組み合わせた翻訳サービスについて、その開発の経緯と現状を紹介しました。また、マイクロソフトの Chris Wendtは、同社が提供する Collaborative Machine Translationについて説明し、利用者が、自分のデータでマイクロソフトの SMT エンジンをトレーニングしながら

利用し、また Microsoft のシステムはユーザーから提供されたデータを利用してさらに品質を向上させていく、というサイクルの可能性について語りました。

評価システム

MT の利用や研究における永遠の課題である評価システムについても発表がありました。TAUS の Rahzeb Choudhury 氏が、従来の one size fits all 的な評価システムでは実現できなかった、コンテンツのタイプや機能、利用者側のニーズ、テキストの寿命などを考慮して適切な評価基準を選択する、動的かつカスタマーの満足度を重視した MT 品質評価の枠組みを確立するためのイニシアチブ「TAUS Dynamic Quality Framework」を紹介しました。

まとめ

以上、多数の発表のなかからいくつかの内容について簡単にまとめさせていただきました。全体的には、どの発表も非常に現実に即した有用な内容だったという印象です。また 2 日とも会場近くのレストランでブッフェ形式の昼食が用意されており、初日の夜には Networking Dinner も開催されるなど、参加者どうしが交流する機会が豊富に用意されていました。このような場を利用して、プログラム中の質疑応答ではなかなか踏み込めないような詳細な事柄についてさまざまな人と直接対話するチャンスが得られたことも大きな収穫でした。

Report

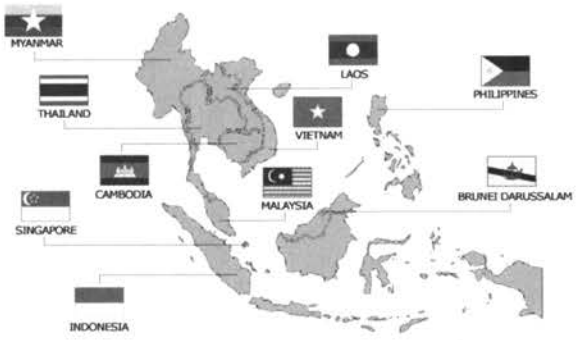
A Current Status of Machine Translation in Thailand
~Network based Machine Translation~

Thepchai Supnithi and Chai Wutiwiwatchai
National Electronics and Computer Technology Center, NECTEC, Thailand

Thailand has a long history of machine translation, started since 1980. There are a lot of approaches on machine translations, rule based, example based and statistical based machine translation which are ongoing research. A lot of resources are developed for the Thai NLP standard. There are two main resource related projects. The first project is Thai national corpus (TNC) project, a national project which aims to develop a 80-million word corpus for Thai language, by Department of Linguistics, Chulalongkorn University [1]. This project collects a lot of raw resources from a lot of publishing companies. The second project is A Benchmark for Enhancing Standard for Thai Language Processing project (BEST) by National Electronics and Computer Technology Center [2]. This project launched a 8-million word segmented corpus called BEST corpus. The objective of this corpus is to develop a Thai standard word segmented corpus. Best corpus was used in Thai word segmentation contest since 2009 both internal and international episode and make Thai word segmentation technique significantly improve. Based on BEST corpus, a lot of Thai resources such as, POS tagged corpus, Name entity corpus, tree bank are developing. All of resources are aimed to hybridize to current MT approaches to improve the accuracy.

Research networking becomes an important issue for machine translation. Recently, we concentrated on how to communicate with other countries for the collaboration. Network based machine translation is focused to translate across languages. There are two projects that are important issues in Thailand.

The first project is network-based ASEAN machine translation project [3]. As you may know that ASEAN Economics Community (AEC) is announced and will start to implement in 2015. ASEAN will become a single market with free flow of goods, investment, capital and skilled labor. Language barrier is one of the major issues, since most of members countries are not skillful in English although English is set to be a common language in ASEAN. All ten ASEAN countries agree to join networked based text-to-text machine translation development project for nine languages, Brunei and Malaysia for Bahasa Melayu, Cambodia for Khmer, Indonesia for Bahasa Indonesia, Laos PDR for Lao, Myanmar for Myanmar, Philippines for Filipino, Singapore for Chinese, Thailand for Thai and Vietnam for Vietnamese under the support from ASEAN Committee on Science and Technology. This project is aimed at developing a public service on tourism domain.



ASEAN Countries member list

The kick-off meeting on The Network-based ASEAN Languages Translation

The second project is networked based speech-to-speech machine translation project [5]. Thailand participated in U-STAR project to promote English ↔ Thai speech-to-speech machine translation. Currently, there are 26 institute from 23 countries join this project. They collaboratively developed the multilingual speech translation system to provide the translation service via publicly-released client application, by connecting the servers of U-STAR member institutes mutually, with the communication protocol which was implemented based on the ITU-T recommendations F.745 and H.625. We also joined to release the speech translation application, "VoiceTra4U-M" for iPhone, and launched a global field experiment. The first trial are implemented for the London Olympic games which is coming up in July.

With the network based machine translation approach, we aim at developing a practical machine translation as a public service in various platforms, both text translation and speech-to-speech translation. Machine translation for specific domain are considered, currently we are focusing on tourism, health case, patent and law domain.

U-STAR member list

U-STAR meeting for London Olympics 2012

Reference

[1] TNC: http://ling.arts.chula.ac.th/tnc2/ (In Thai)
[2] InterBEST:
http://thailang.nectec.or.th/interbest/index.php?option=com_frontpage&Itemid=1
[3] BEST: http://thailang.nectec.or.th/best/ (in Thai)
[4] ASEN-MT: http://www.thaimt.org/aseanmt/index.php
[5] U-STAR : http://www.ustar-consortium.com/

レポート

（再掲載）

機械翻訳の実用とポストエディット

日本オラクル株式会社　斎藤　玲子

はじめに

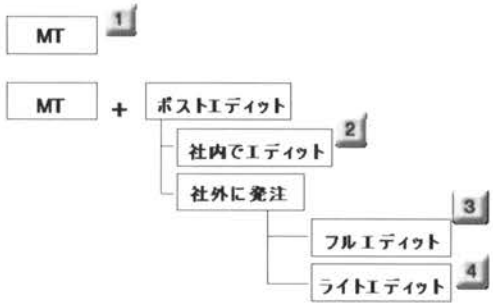
翻訳を発注する立場から見た、機械翻訳導入のメリットは何でしょうか？コスト削減、納期短縮、人手不足の解消 — まるで魔法のように、すべてを解決してくれる素晴らしい存在であって欲しいと、期待は膨らみます。しかし、現実問題として、そこまでの道のりは平坦ではなさそうです。困ったときに打つ手は 3 つあります。

1. 環境が整うのを待って、耐える。
2. 今あるもので何とかする。
3. 新しい手段を探して投資する。

ここでは、2 の「今あるもので何とかする」方法で、現段階の機械翻訳を実際のビジネスにどうやって生かしていくことができるか、機械翻訳を導入する側の視点と作業する側の視点から考察してみたいと思います。

機械翻訳の 4 つのオプション

機械翻訳を使う場合、まず機械翻訳の結果をそのまま使う、機械翻訳した結果をポストエディットする、という 2 つの選択肢があります。後者の場合、ポストエディットを社内の人材を使って行うのか、社外に発注するのかで道は分かれます、社外に発注する場合は、さらにライトエディットをするのか、フルエディットをするのかに分かれます。つまり、以下のようなツリー構造になります。



「フルエディット」と「ライトエディット」という言葉について説明します。「フルエディット」は「正確で読みやすい内容に編集する」ことで、「ライトエディット」は、「正確な内容に編集する」ことを指します。どのように違うのか、例を挙げてみます。

英語: View the different location groups to which a location is associated.

MT: ロケーションが異なる場所に関連するグループを表示

ライト: ロケーションが関連する、異なるロケーショングループを表示

フル: ロケーションが関連付けられている、さまざまなロケーショングループを表示

ライトエディットでは、語順を入れ替えて読点を追加しただけです。フルエディットでは、それに追加して associated と different の訳を変更しています。フルエディットでは、ゼロから翻訳する場合と結果はほとんど変わりません。わかりやすい反面、時間の節約についてはあまり期待できません。

社内でエディットする場合は、基本的に「ライトエディット」になるだろうと想定しています。これは、作業にあたるのが選任の翻訳者ではなく、翻訳される内容の分野で実務に携わっている人であることが多いためです。

上記をふまえて、4 つのオプションをスピード、品質、コストの 3 点から比較し、それぞれに適した用途を考察してみます。

1. MT のみの場合

	1 MT	2 社内ライト	3 社外ライト	4 社外フル
スピード	◎	△	○	△
品質	△	○	○	◎
コスト	◎	◎	○	△

1 つめの MT のみでは、スピードとコストはベストです。一方、まだ品質については難しい面があります。あるテストで Moses というオープンソースの統計ベースの機械翻訳エンジンの品質評価をしてみると、現在は「そのままでは、半分ぐらいは理解できる」レベルという結果になりました。このオプションに適しているのは「オンラインで大量で無償で提供するもの」で、提供方法に工夫が必要です。「これは機械翻訳です」などと、ユーザーに事前に知らせ、元の英語が参照できるようにリンクを設けておくなどして、ユーザーが「英語でしか情報がないよりは便利」と思ってもらえることを目指します。

2. MT + 社内でエディットする場合

	1 MT	2 社内ライト	3 社外ライト	4 社外フル
スピード	◎	△	○	△
品質	△	○	○	◎
コスト	◎	◎	○	△

2 つめの MT+ 社内でライトエディットの場合、品質は「正確な翻訳」レベルになり、コストも外注費がかからないという点では◎です。一方、スピードについては、社内で選任でない人が作業にあたるため、そのトレーニングやスケジュールを考えると、メリットがあるとは言えません。このオプションに適しているのは、1 件あたり 10,000 words 前後と量が少なめで、内容は作業者が専門とするものです。テクニカルノートなどの技術情報が一例です。

3. MT + 社外でライトエディットする場合

	1 MT	2 社内ライト	3 社外ライト	4 社外フル
スピード	◎	△	○	△
品質	△	○	○	◎
コスト	◎	◎	○	△

3 つめの MT+ 社外でライトエディットの場合、すべてについて合格点であると言えます。外注費はかかりますが、通常の翻訳費用よりは低く抑えられ、品質は「正確で理解できる内容」になります。作業量が通常翻訳より少ない分、スピードも上がります。このオプションに適しているのは、「定型的な文章の多い、大量でオンラインのドキュメント類」です。技術分野でいえば、管理ガイドやリファレンスマニュアルなどが一例です。

4. MT + 社外でフルエディットする場合

	1 MT	2 社内ライト	3 社外ライト	4 社外フル
スピード	◎	△	○	△
品質	△	○	○	◎
コスト	◎	◎	○	△

4 つめの MT + 社外でフルエデットの場合、品質はゼロからの人による翻訳とほぼ変わりません。スピードとコストも、人による翻訳と同じぐらいかかることが多いものです。この 2 点を改善するには、分野に特化した MT エンジンへの教育、編集環境の効率化、および編集方法の効率化が必要です。それらが改善されれば、オンラインだけでなく、紙で出版されるドキュメントや、わかりやすさを求められるものにも、正確でわかりやすい内容の翻訳を割安なコストでより早く提供できることになります。

編集方法の効率化

上記で挙げた 3 点の改善方法うち、「今すぐ何とかできるもの」は何でしょうか。エンジンの教育と編集環境には特殊なツールや環境および時間が必要です。そこで、ガイドラインなどのソフト面で対

応できる「編集方法の効率化」に着目してみます。

ポストエディットは、「翻訳し直し」ではありません。かといって、「どんなに悪い翻訳でも、修正して使わなければならない」わけでもありません。では、どのようなガイドラインがあればいいのでしょうか。

ポストエディットのガイドライン

ガイドラインの有無で作業の効率は大きく変わります。ガイドライン作成の前後で効率が 2 倍になったケースもあります。

ガイドラインには「目標」と「手法」の 2 点が必要です。つまり、「どのような品質にするのか」と「どのようにして編集するのか」です。品質については、「正確でわかりやすい翻訳」なのか「正確な翻訳」なのかを特定します。もちろんこれは、作業者だけが目指しても仕方ないので、発注する側との合意が事前に必要になります。具体例を挙げて、作業する前に相互で確認しておきます。正確な翻訳を目指すのであれば「スタイルの不統一や読みやすさは求めない」ことを明確にするべきです。

「手法」については、どうすればスピードを上げることができるかに着目し、具体的に方法を提案することが重要です。私が気をつけている点は以下です。

心構え

1. 「MT は正しい翻訳を出力するべき」という考えから離れる。
2. 「使えるものは使おう」という精神でのぞむ。
3. それでも「自分で翻訳したほうがはやい」と思えば、そうする。

「心構え」をするのは、MT に対して中立的な立場を保つためです。MT に対してネガティブな気持ちのある人ほど、効率は下がる傾向があります。

実作業

1. 日本語を読み、英語を読む
2. 短い時間で「理解できる内容になるか」チェックする。
3. Yes なら編集、No なら再利用またはゼロから翻訳する。

実作業の 1 と 2 をすばやく行うことで、ゼロから翻訳するだけよりも、少しでも効率を上げられるようになります。3 をより具体化すると次のようになります。

1. MT を修正して使う
2. MT を参考にする
3. MT を使わない

2 の「参考にする」とは、自分で翻訳するが、その中に MT にある表現を使うということです。さらに、1、2、3 の使い分けの目安も示すとよいでしょう。たとえば、以下のようにします。

作業ガイドラインの例

1. 以下の場合は MT を修正して使う
   - 1.1 一度読むだけで 80% 以上が理解可能
   - 1.2 修正箇所を決めるために日英を比較する回数は 1 回
   - 1.3 必要な編集作業は 2 回まで
2. 以下の場合は MT を参考にする
   - 2.1 一度読むだけで半分強が理解可能
   - 2.2 参考にする箇所を決めるために日英を比較する回数は 1 回
3. 以下の場合は MT を使わない
   - 3.1 一度読むだけで半分以上が理解不能
   - 3.2 参考にする箇所を決めるために日英を比較する回数が 2 回以上

まとめ

機械翻訳を発注する側としては、MT の使い方(オプション) とそれぞれに適した翻訳対象を見極め、期待する品質を決めて、作業者に伝えることが重要です。最初は「正確な翻訳」で良いと思っていても、出来上がったものを見ると「正確でわかりやすい翻訳」に修正したくなることがあります。しかしそれでは、低コストと短納期で通常翻訳を求めて

いることになります。このようなことが起こらないように、作業を開始する前に、発注側と作業側で目標とする品質について丁寧に共有し、合意しておくことが大切です。作業側も、目指す「品質」の具体例を挙げ、積極的に発注側に提示するべきです。そしてその内容を「作業方法」とともに、ガイドラインとして作業者間で共有することが、ポストエディットの品質の一定化と作業の効率化につながると考えます。

【お詫び】

前号（P.59～62）にて掲載させて頂いた斎藤玲子様の本報告につきまして、表中の記号に誤記があり、斎藤様のご指示とは異なったままでの掲載となりました。このため、今号での再掲載を以って訂正させて頂きますとともに、深くお詫び申し上げる次第です。

Report

Pangeanic's Do-It-Yourself Machine Translation: User Empowerment and User-Driven MT Processing

[Pangeanic の DIY 機械翻訳：管理権限をユーザーに委ねた ユーザー主導型機械翻訳の全貌]

Elia Yuste, Manuel Herranz, & Alexandre Helle
eyuste/mherranz/ahelle@pangea.com.mt

Pangeanic / B.I Europa
PangeaMT Technologies Division
Valencia, Spain

A.-L. Lagarda, M. García, J. Pla-Civera, M. Blasco, A. Morellá, & J. Mallach
alagarda@iti.upv.es
Institute of Computer Technology (ITI)
Universidad Politécnica de Valencia (UPV)
Valencia, Spain

1. Introduction

This paper reports on how Pangeanic's machine translation (MT) offering, PangeaMT[1], has evolved to include technical components that allow the user to perform MT-related actions that were in the exclusive hands of the MT provider in the past, namely engine creation and updating or retraining. We explain why we thought this was a necessity among our user-base in the translation industry and enterprise market; in other words, why and how we decided to set the DIY[2] footprint in the translation automation arena. The components of a **PangeaMT DIY** solution as released in late Spring 2011 are presented. This information is framed taking into consideration the full picture of PangeaMT technologies and their business models and services until 2012/Q1. Given PangeaMT's innovation-driven slant, we then revamped the concept of MT-DIYing and decided to make it not only available for PangeaMT DIY self-hosted solution clients, like in 2011, but also for users working in the so-called new **SaaS Power** mode. This is a/n r/evolutionary MT SaaS[3] type, whereby users not only enjoy the typical PangeaMT benefit of machine translating as much and as often as needed but they are empowered to retrain their engines online and fuzz-free. This represents a major breakthrough in the translation industry, particularly in the case of organizations and language service providers (LSP) that are really keen to be in full command of all their MT processes, not only machine translating, in a transparent and efficient fashion but yet cannot afford the time or the little to medium range investment to host a DIY solution themselves. Nonetheless, those multilingual content production agents, handling highly confidential data, will continue to be interested in self-hosting their MT solution and growing their own MT ecosystem. Having them in mind, the PangeaMT DIY offering first released in 2011 can be delivered in an integrative, intuitive and user-layered platform called **PangeaMT Full Power** since June 2012. This article concludes with a reflection on how we think user intervention leads to user empowerment, an increased intake of MT technology and, ultimately, to overall MT technology advancement.

---

1 www.pangea.com.mt

2 DIY stands for Do-It-Yourself. In this context, it refers to empowering the user of a machine translation (MT) system to go beyond the activity of an MT query or request via a command line or web panel, and be able to manage engine training data and engines, updating them when need be, without having to resort to the MT provider.

3 SaaS stands for Software as a Service.

2. Birth of the PangeaMT DIY Concept

Pangeanic/B.I Europa (Pangeanic for short, an associate member of B.I Corporation in Japan) is a Spain-based language service provider (LSP) with close operational links in Asia and a machine translation division called PangeaMT. In Yuste et al. (2011) we provide an introduction on Pangeanic's transformation from a translation agency into a global LSP and a statistical machine translation solution customizer. This is worth reading if you are interested in finding out more about the early days and the philosophy of PangeaMT and its evolved connection to Moses[4], i.e. PangeaMT has built in several modules and pre- / post-processing scripts on top of Moses, as well as interfaces and retraining procedures in order to develop an industry-tested alternative[5] to off-the-shelf, more rigid MT products.

The fact that Pangeanic wears two hats, that of an LSP and of an MT provider put the company in a privileged position from the very beginning to contemplate different MT technology deployment scenarios. We listened not only to our internal users, i.e. translators, reviewers/QA checkers and project managers, but also to our clients, who interestingly enough come both from the enterprise market and the international language service industry. While the former usually play the role of buyers and hardly own a full localization department with a link to MT or other HLT[6] areas (unless we talk about cutting-edge IT and Internet related players who may even have experience in MT development), the latter feel the urge to adopt MT technology in their processes but, depending on the LSP size and background, they may still be somewhat confused as to the MT paradigm or product that allows them to translate more for less.

While multilingual enterprises are deploying MT for internal communication and information dissemination, and lately even in combination with business intelligence analytical tasks, particularly of rather more volatile content such as UGC[7], the LSP primarily looks at MT as a multilingual technical content production-enhancement tool and a means to react against the ever-increasing demand for publication quality translations in squeezed TAT[8] and drastically constrained budgeting conditions. Therefore, what motivates these two types of MT user/customer segments is cost-effectiveness and ROI[9] even if their deployment scenarios and goals may sometimes differ.

Another aspect they have in common is a demand for rapid and user-focused system's performance, and this may go well beyond the actual step of machine translating. Enterprise market buyers are keen to elicit an MT request from a web interface or an API for integration in other applications, and this has to be run as quickly and smoothly as possible. If they have an in-house localization department, they will usually perform a black-box evaluation of outputs across their languages of use and from several MT developers until they make up their mind about what MT provider(s) to choose. Real engine customization, previous in-production experience in the same specialization domain, scalability and robustness will be determining factors for success. Moreover, buyers with a keen interest in MT and a view to intervene in MT processes, although not willing to reinvent the wheel

---

4 www.statmt.org/moses

5 See Yuste et al. (2010) for more information.

6 HLT stands for Human Language Technology.

7 UGC stands for user-generated content.

8 TAT stands for Turn Around Times.

9 ROI stands for Return On Investment.

or adapt the Moses toolkit to their needs, are becoming really appreciative of MT solutions that allow for **rapid engine customization** and in particular, **routines for updating engines at will**.

LSPs have been confronted with the challenge of using MT without having technical/computational linguistics background and, on many occasions, the resources to invest in MT technology self-development. However, a growing number of LSPs are getting bored of being just 'MT output-only handlers.' Instead, as it happens with MT users from the buyers' segment, they are becoming more and more interested in learning about MT and playing a role in MT processes, such as engine updating, particularly if this is made easy to them. Even in the case of LSPs that were more experienced with MT, we noticed a clear dependence from the MT provider. This reminded us of Pangeanic's own dependence from our programmers whenever we felt that an engine needed updating or retraining.

With a view to fostering even more accessibility to **flexible and customizable MT solutions** as well as further **independence from MT providers for engine experimentation and retraining**, Pangeanic officially released the **PangeaMT SMT DIY** Solution for E-FIGS[10] right before the *Localization World Conference* in Barcelona in June 2011. Its main asset consisted of a trouble-free automatic engine training routine, so there was no need to knock on the MT provider's door any more and also pay for every engine updating. Since then PangeaMT DIY, being a self-hosted solution, has proved to be appealing to enterprise and LSP MT users alike, particularly in deployment scenarios where **data confidentiality** is at stake.

3. PangeaMT DIY as released in 2011

PangeaMT SMT DIY included a **training data set structure**, similar to that of an FTP window, where users would simply leave their bilingual aligned files in TMX 1.4b format. Thanks to their associated automatic training routines, engine training and updating could be done automatically, either on a time basis or incrementally (after 50Mb, 100Mb, 300Mb of additional data are detected – a setting that could be determined by the client).

This was paired with a **Control Panel** that showed an engine listing table, with a focus on engine status and automatic quality metrics, such as BLEU for the direct and reverse translation directions, as shown in Fig. 1. PangeaMT DIY was then the first solution on the market to incorporate informative engine status charts with quantitative and qualitative data available on a 24/7 basis.

[10] English into/from French, Italian, German, and Spanish. Offering first this language bundle, then some combinations thereof, other languages, and linguistic families, according to clients' specifications and needs (e.g. Brazilian/Iberian Portuguese, Scandinavian languages, etc.).

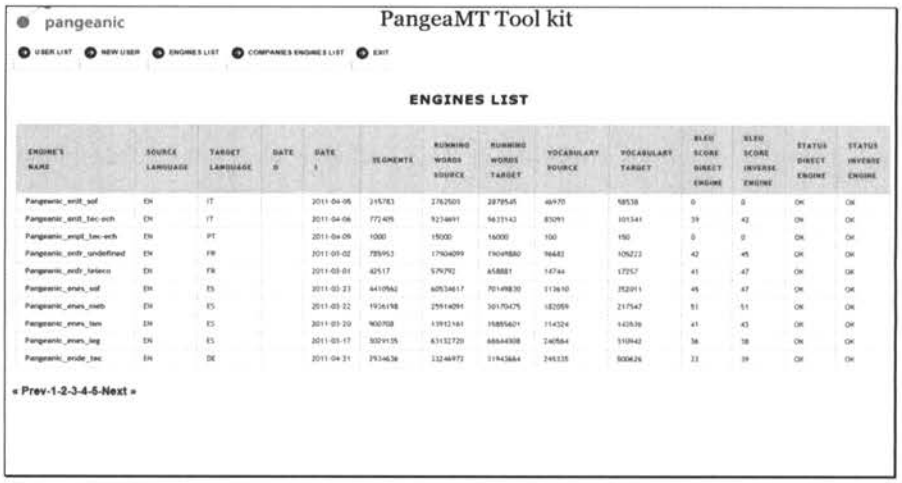


Fig. 1 *Control Panel: Example of Engine List with Statistics*

No computational linguistics knowledge was necessary to navigate through the Control Panel or to perform an engine training task. For instance, just by knowing for sure which bilingual data assets or TMs to throw in a given structure data folder, a new engine or an existing engine could undergo a training process - as simple as that. As soon as the (re)training was completed, the statistical information about the engine would be made available on the Control Panel's Engine List section. No other MT offering had encompassed automatic training routines while displaying their effects on the engines immediately.

The Control Panel also had a basic user activity tracking display and allowed for unlimited users' creation. Any newly created or updated engine could then be assigned to a certain user or company's section in an intuitive application within the Control Panel, just like this:

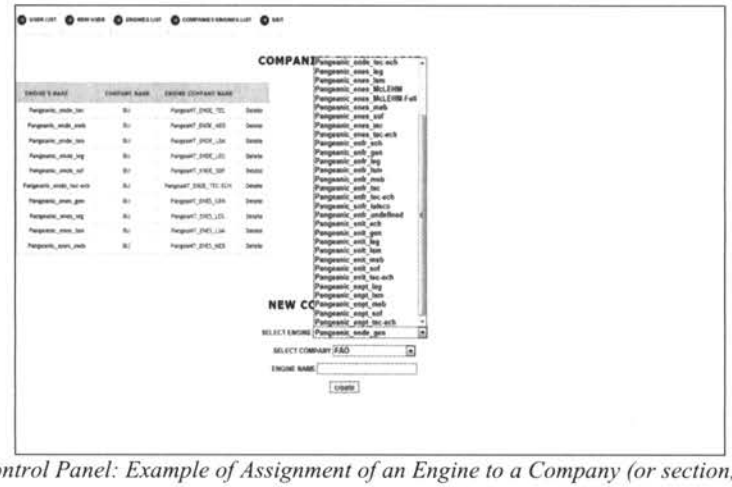


Fig. 2 *Control Panel: Example of Assignment of an Engine to a Company (or section, department)*

As far as MT requesting and user-application interaction are concerned, there was an intuitive Google-like **web-based translation panel** and, most interesting, a **%MT file translation feature** whereby pre-translated XLIFF

or SDL Trados TTX files could be translated after a certain match percent had been leveraged[11] from the translation memory by the user. This is fully customizable by project, so one chooses exactly at what pre-translation level MT will come in, not touching any material leveraged from the TM. Other useful feature available in the translation panel was the **Glossary** TXT file upload that helped fix terminology and preferred expressions, such as *untranslatables* or DNTs[12] in the background prior to the MT output – by the way, quite a historic request to SMT developers.

3.1 Perception of PangeaMT DIY Technology in the industry: Benefits and *side effects*?

MT competing players that are not willing to be as open-minded and **user-empowering** as Pangeanic prefer to have an immovably enshrined focus on providing solutions that are output-driven[13] and not **tool-driven**. Coming also from the Language Service Provision sphere, we at Pangeanic consider that this may be counterproductive for the industry. Allowing for new MT user-focused paradigms and easing typically user-obscure MT processes, enabling interested users to retrain engines or reconvert their TM assets into *fuel* for their engines, can only lead to a wider and more motivated adoption of the technology. Doing this does not mean though that Pangeanic is offering DIY MT tools without taking care of the necessary steps and tasks that go in a real MT customization, as some may have tried to argue.

On the contrary, Pangeanic also offers **Data Consultancy** as part and parcel of every first-time PangeaMT custom engine creation, independent from the mode in which it will be made available to the user: in a PangeaMT SaaS mode or as a self-hosted solution, or from whether the user will be just a translator or an organization interested in performing DIY tasks, too. Clients, whose existing bilingual data may be unclean or scarce for engine training purposes, are particularly appreciative of this type of service. Depending on the condition and size of the starting training corpus, a customization may then encompass the training of one than more engines, with the project resulting in at least two or three engines, ranging from one trained with the client's TM material only, plus others having been trained with other suitable datasets.

When a PangeaMT DIY customer gets hold of their solution, much effort has thus gone into analyzing their training data, seasoning it with any necessary advanced filtering/cleaning[14] pre-processing techniques and

---

11 PangeaMT's TMX / XLIFF /TTX parsers facilitate the user's preference for any CAT tool as post-editing environment (a TMX or XLIFF editor can be used, and even Tag Editor, which will identify pre-translations as coming from MT! and stop at each segment. This allows users to leverage their TMs as well.

12 DNT is a localization industry acronym that stands for "do-not-translate" a given word or expression, usually due to product marketing and corporate identity related issues. A typical example is the brand name Apple. For a computational linguist, a DNT is somewhat similar to the NLP concept of *named entity*.

13 These MT companies rather stick to traditional word-based pricing business models, imposing limits about the number or weight of MT translated word on the user.

14 Automatic cleaning routines in the PangeaMT pre-processing module that may be applied for automatically detecting suspicious or unclean translation units and extracting them from the training corpus, that is, the bilingual files or TMX files that the client provides the PangeaMT team for a custom engine creation. If interested, the extracted suspicious units can be handed in for later verification by the user. Suspicious units are those which contain spelling errors or punctuation/numerical inconsistencies, or in which the source language segment is very similar or is identical to that of the target language. The client commissioning the custom solution should be aware of the fact that applying these filtering methods has a pruning, or cleaning effect, therefore leaving their data clean and free of segments that may represent a hurdle to a statistical based translation output. However, the clean training set usually gets smaller in size.

additional translational assets from public and reliable sources[15] and repositories. We regard this as a kind of necessary *planting-the-seeds* operation to ensure that if the client later opts to work on their own in a PangeaMT DIY framework, they can then operate intuitively and freely, yet harvesting more than adequate MT offspring, in terms of new and self-retrained engines and their output. It is only then that true MT DIYing can come into play, with minimum intervention from Pangeanic as a MT provider. Yet the PangeaMT team remains at the client's disposal for support or training purposes at any time after the PangeaMT DIY solution has been made available.
The essence of MT DIYing, also as initiated by Pangeanic, has been discussed in detail in the article put together by Simpkins (2012), which targets a readership interested in the hottest localization industry topics.

4. Going Full Power – The PangeaMT DIY Concept Gets Revamped in Spring'12

MT DIY features of engine self-training and updating as well as round-the-clock engine status information access were at the core of the so-called PangeaMT SMT DIY offering launched in Localization World Barcelona in June 2011. These DIY components were easy to use and truly well-received by testers and clients. However, we wondered if we could even go one step forward and make all these DIY features accessible from a single spot. This would in principle facilitate and speed up remote installation in the case of self-hosted solutions. PangeaMT DIY components would have to remain fully functional but also further interlinked and viewable from a platform, which would become highly versatile on the basis of user profiles, that is, the ones using it.
The new **PangeaMT Platform** released in 2012/Q2 is the result of listening to our own needs as an MT-proficient LSP and those of our MT clients with a strong desire to implement customized MT without MT request limits, and with full tracking of all MT process actions and top-notch DIY capabilities – all without having to log in or open several different applications!
Most importantly, engine training and updating will also be available to users under the **SaaS Power** framework. This is definitely Pangeanic's 2012 innovation with regard to PangeaMT, as discussed in the next section. All in all, users demanding the PangeaMT DIY solution as borne in 2011 are now happy to find even more features that guarantee their independence from the MT provider and a full control over engines and data, hence the new name **PangeaMT Full Power**.
Integrative in nature – all components that made up the PangeaMT DIY offering of 2011 are made available from a sort of one-stop spot, that is, a password-protected, web-based, unified interface with enhanced MT process informative functionalities that offer more or less options on the basis of a clear-cut user profiling.
Let us know explore those different profiles from a bottom-up approach. Fig. 3 shows the landing page of someone logged in as a translator in the new PangeaMT platform, where no further advanced features are offered. This would be the platform's basic or level 1 user profile.

[15] An example of this is TAUS – see www.translationautomation.com/

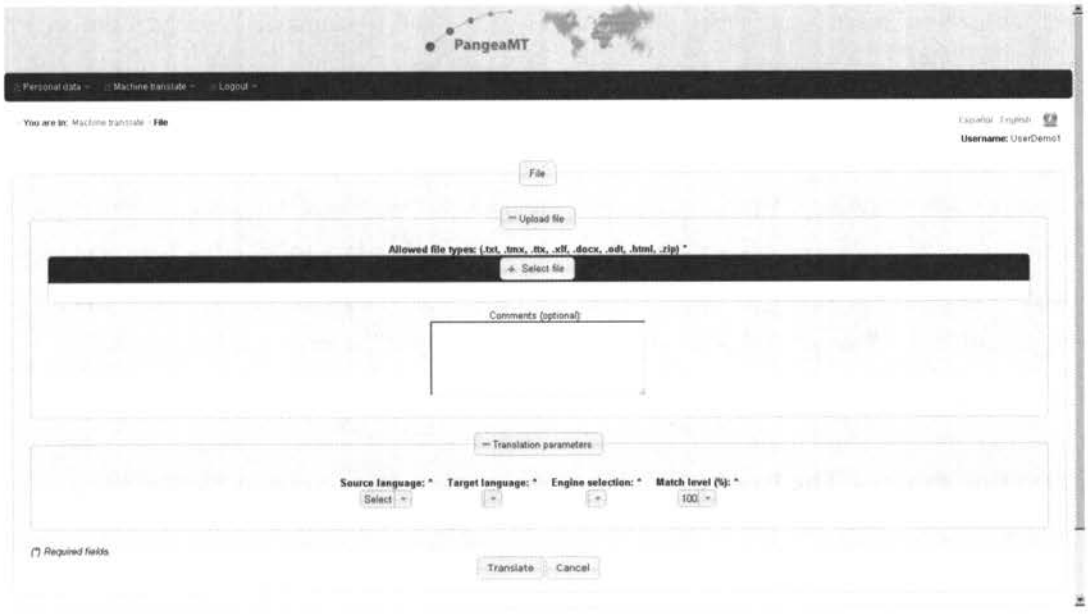


Fig. 3 *Translator-only Profile: Detail of Machine Translate > File Upload function window view*

In particular, this figure shows how to upload a file or a number of files (subsequently or in an archived .zip file) to request machine translating. Worth noting are the Match level (%) function as one of the required translation parameters as well as the possibility to upload a Glossary (further down in the same window) at the time of requesting MT. The translator will be able to make as many MT requests or upload as many files for MT as necessary, without any limit whatsoever.

Every translator's activity gets logged in the so-called **Translation report**, available from the Machine translate pull-down menu as well. The translator can then query about and get access to any MT jobs of them by specifying a given timeframe. These jobs exclusively and not the ones requested by any other user get listed in inversed chronological order. This is why our translator does not get to see any MT jobs from the selected JP into EN engine, as s/he has not performed any MT request with this engine yet. Às shown in Fig. 4, there are no results to see here. The Translation report functionality can be really useful, as all kinds of administrative information and the machine translated files from the jobs listed are accessible here, too. The person logged as company's administrator, enjoying advanced DIY features and unrestricted data access rights, will be the only one accessing every translator's job and related info.

Fig. 4 *Translator-only Profile: Detail of Translation Report – No results found (no jobs requested by this user with this engine yet)*

The next level user profile, level 2, could be appropriate to a project manager (PM) interested in having the same system's functionalities as the translator but also more administrative information about the custom engines belonging to the company. This user profile in the new platform is benefitting from the same engine control information available in the Control Panel available within the first PangeaMT DIY suite of 2011 (see previous section, particularly Fig. 1). This information is now accessible through the so-called **Memories, Domains, Engines** pull-down menu, particularly in the Engines section, which is the only one partly available for this type of user. When going to Memories, Domains, Engines > Engines, the user sees the engine chart shown in Fig. 5. Apart from some admin info, such as *creation date* or *last training date*, the PM gets to know about the status of an engine, for instance, whether it is *Active*, *Inactive*, in *Training*, etc. This is essential for day-to-day translation project management purposes.

When the PM is keen to know more about a particular engine, clicking on the *+info* button will lead to a really informative page, containing all statistical and administrative details about the various versions of that engine. See an extract of this information in Fig. 6 on next page. Right at the bottom of the displayed information, the user can read about all the bilingual files making up the Corpus List, that is, the files that have served as a training set for that engine version. Any corpus file can get selected and accessed for a quick view, if interested. The same transparency principle applies to the engine training process itself. Should the user be keen to know more about how this happened, there is an engine training log file for both directions of the engine version, direct and inverse[16].

[16] A into B language direction and the opposite one. PangeaMT engines are always bi-directional, containing in fact two training models corresponding to the two translation directions, direct and inverse.

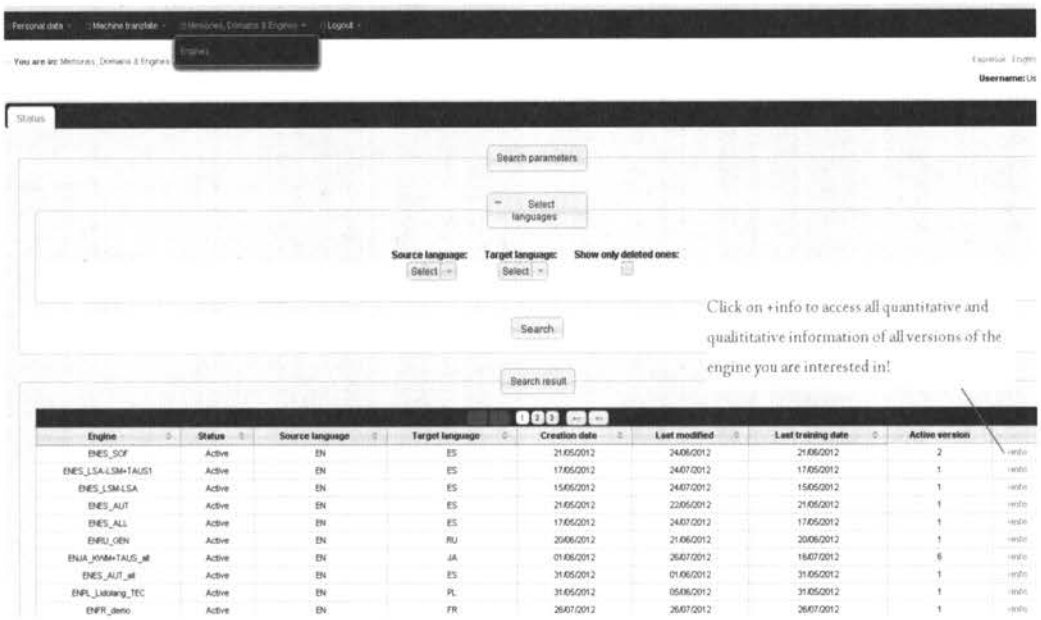


Fig. 5 *Memories, Domains & Engines – Engine List: Status (as seen by a level 2 user profile, e.g. a PM)*

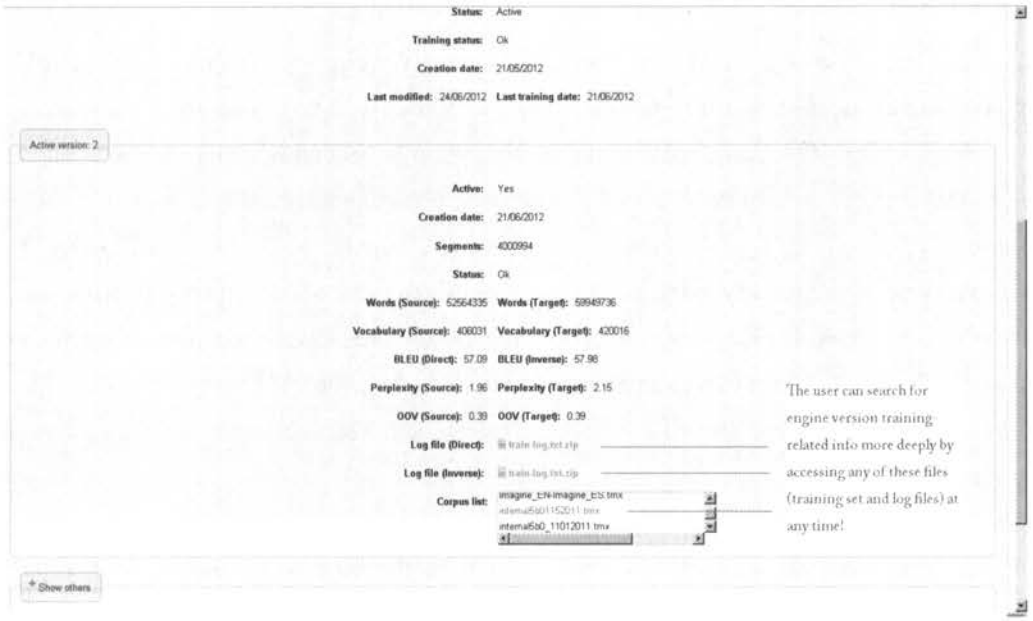


Fig. 6 *+info: Detail of Qualitative and Quantitative Info about an engine version of interest*

The information displayed here gets updated constantly and is available on a 24/7 basis. Fig. 6 shows a detail of statistical information about an engine version, containing e.g. the BLEU[17] scores of both corresponding translation models. The page can display this type of information for all versions of an engine at the user's will. This is particularly of use if a PM wishes to know how a translation engine has improved across the different versions, that is, after one or several retrainings, just by looking at the BLEU figures along the different engine

[17] Information resulting from other qualitative metrics may be included as per client's request. BLEU stands for Bilingual Evaluation Understudy. An explanation of BLEU is included in the PangeaMT Platform's online help. For further details, please refer to Papineni (2002) or read the entry on Wikipedia (en.wikipedia.org/wiki/BLEU).

versions. In particular, this represents an enhancement with regard to the first PangeaMT DIY solution, where the Control Panel was capable of showing this kind of information but only about the last active version.

An advanced/more senior translation PM or the person(s) designated in the company to act as MT manager needs to have a more advanced user profile in the PangeaMT platform, i.e. they are the ones who will be able to enjoy real DIY functionalities. Logging onto the system as this so-to-speak level 3 user profile, the **Memories, Domains & Engines** pull-down menu is now fully active in all its sections.

Let us now explore what this user profile can do! The **Memories** subsection has two tabs available: *Search* and *Upload*. Intuitively enough, the *Search* tab is resorted to in order to search for any translation memories (TMs) that have been previously uploaded in the PangeaMT data repository, which contains all bilingual translation assets in **TMX 1.4b** format and ready to use for engine training purposes. The search resulting window will display the memories the user is after according to the search parameters previously selected, such as language direction, uploading user, or time frame. The list of displayed memories includes administrative and statistical information, and, as it happens in the case of the Engine Status section discussed above, the user can access the files themselves, by clicking on the Filename highlighted in orange. At the end of every TM line in the table, there is a clickable button called **Edit**, which not only provides extensive information on the TM file itself but allows for a fairly powerful functionality, namely that of associating this TM file to an existing domain. This drag-and-drop easy operation comes handy if our advanced user thinks that it makes sense that this TM file, belonging to X domain, gets also activated in another domain.

The **Memories** *Upload* tab allows the user to upload as many memories as they see fit. This can be done one by one or in a zip archived file. The most important thing is that the user remembers that the system should hold all TMs in the TMX 1.4b standard format. Fig. 7 shows the looks of this tab. As soon as the memory asset(s) get(s) uploaded in the system, this new material is searchable from the Memories *Search* tab described above.

The next section in the **Memories, Domains & Engines** pull-down menu is **Domains**. Natural Language Processing (NLP) systems and applications that are specifically created in or adapted to a restricted domain, that is, a specialized area of knowledge, tend to perform better than those handling or processing open-domain content and theoretically speaking at least, they are less dependent on a vast amount of data. Modeling and training in the PangeaMT framework, as opposed to well-known generalist counterparts such as Google Translate or Bing, has focused on restricted domains as the company itself has a language service provision tradition to serve industry vertical clients that are highly specialized.

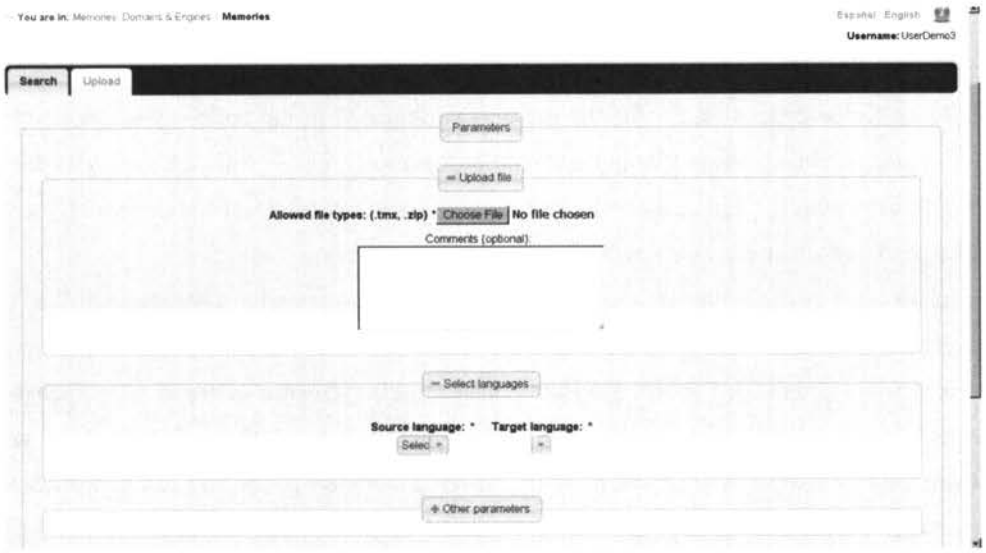


Fig. 7 *Detail of Memories, Domains & Engines > Memories > Upload*

Therefore, in our context, the concept of *domain* is always intertwined with a given industry sector, such as automotive, banking, biotechnology, tourism or renewable energies, to name a few, and the corporate language specificity[18] of the client commissioning the custom MT solution. PangeaMT custom solutions are thus in-domain and client-specific. Some of the domains, meaning therefore *client-driven domains*, we have tackled pose inherent challenges, such as data scarcity or training data showing text types that are closer to natural language than controlled, technical language. This is one of the reasons why the notion of domain in PangeaMT had to be flexible, in that we realized that some customizations, if not aiming to be completely open-domain, would necessarily have to be kind of mixed-domain and even contain supporting bilingual data coming from trusted sources belonging to a different sector to that of our client. Filtering data as well as picking & mixing data from different domains to experiment and generate better-performing custom engines are common operations. How to facilitate this in an intuitive fashion in the new PangeaMT Platform has been achieved through this menu, whereby memories are held, searchable and uploadable, and then get associated with domains in a flexible and traceable step prior to launching an engine re/training.

The **Memories** *Domain* section has two tabs, *Domain Management* to search for domains already available in the system, and *Domain Creation* to create a new domain, which depending on the user's context, it could well be a sub-domain to designate the content of a corporate division, a product line, etc. An illustrative example of this would be a multi-site pharmacy industry client that purchases the **PangeaMT Full Power** solution hosted at their HQ to machine translate their content in 20+ languages originating from four divisions across the globe. They would manage their TMs and assign them to domains that reflect their four divisions, dealing respectively with nine medical areas, such as cancer, tropical diseases, AIDS/HIV, etc. They could create domains on the basis of their geographical divisions or their medicine coverage areas.

Domains are interpreted somehow more traditionally across our translation agency client-base in that they tend to name their newly created domains using industry sectors names and conventions, such as automotive or AUT.

[18] Mainly in terms of style, language register, terminology including name entities (NE) or untranslatables/DNTs as called in the translation industry, etc.

However, given the system's flexibility to designate domains and engines according to one's needs, the data available and the envisaged scope of the engine, many LSP companies name their domains per client or sector and client, e.g. AUT or AUT_Mitsubishi. This comes handy when the LSP wishes to launch the training of a sort of **supra-engine** of a given domain, encompassing the memories belonging to domains that in reality are client data specifications.

This example leads to the most powerful DIY function of the new platform – that of the **Memories, Domains & Engines** pull-down menu called **Engines**. This section has three tabs, *Status*, *New* and *Training*. While the former two are rather of admin nature, to search for and learn all about the status of all existing engines and to create a new engine respectively, the *Training* tab empowers the user to train an existing or a new engine – that is really the biggest DIY takeaway of the new PangeaMT Platform in its Full Power flavor (or SaaS Power, as we will see next). There are many reasons that may motivate the action of training or retraining an engine. While retraining usually results from the user's wish to expand and improve the output quality of an engine, such ease-of-use to create a new engine opens up lots of possibilities in a day-to-day MT-enhanced localization business.

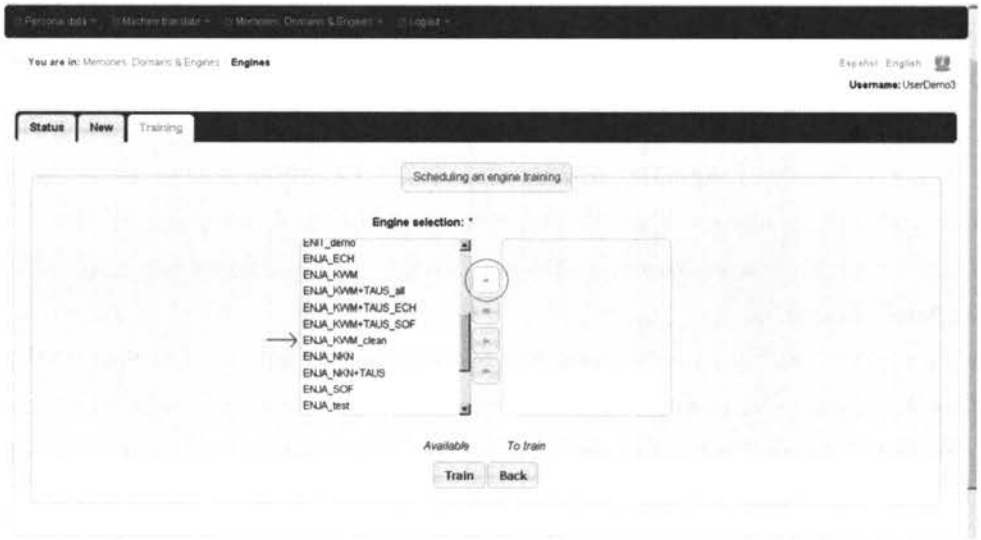


Fig. 8 *Memories, Domains & Engines > Engines >Scheduling the training of a new engine*

Fig. 8 shows the actual moment of dragging and dropping the designation of a new domain, ENJA_KWM_clean, to the right-hand side before pressing the *Train* button that elicits the training of a new engine. ENJA_KWM as well as other mixed-data ENJA_KWM engines existed already. ENJA_KWM_clean now contained the memories provided by the client[19] that had just undergone an advanced filtering (cleaning) pre-processing as part of our ongoing MT customization for this Japanese LSP. In a matter of hours this new engine got ready and automatically viewable to the client, accessing the new PangeaMT platform remotely in a SaaS mode. In just no time they could access their existing engines and this new one, and then assess on their own if this cleaning process had had a positive impact on the output quality by submitting a translation job to the whole range of engines and evaluating results contrastively.

[19] KWM is the short form of a client of ours.

5. DIY Concept Extended: SaaS Power

In the process of revamping the PangeaMT DIY concept, originally thinking mainly of those clients willing to buy a self-hosted PangeaMT DIY solution, we came up with the idea of an integrative platform as explained in Section 4. In parallel to this, we soon realized that this platform should internally be powerful, secure and scalable enough to allow us, as Pangeanic, to bring to market highly innovative SaaS versions of PangeaMT that would not have been conceived without the inception of the platform.

In essence, up to the end of 2011, our existing SaaS clients could already enjoy the benefits of a typical PangeaMT solution and community-oriented client policies: namely, unlimited MT requests, open translation industry standard and interoperability fostering through PangeaMT TMX and XLIFF parsers and generators, competitive opt-out MT SaaS subscriptions including initial training and support, gradual adoption of the PangeaMT technology across language domains through custom engine piloting, and last but not least, total 24/7 engine access and traceability through the Control Panel as described in Section 3 above. However, our PangeaMT SaaS users were still dependent on us, as MT service providers, for some essential MT tasks, such as managing domain-specific and their own training data and retraining engines, which used to come at a cost, even if moderate.

The new PangeaMT platform would ideally have to allow any SaaS user interested in remote engine training to do so online. This has been possible in early 2012/Q2 and is, in our modest opinion, PangeaMT's breakthrough of the year 2012. The so-called **PangeaMT SaaS Power** offers all DIY functionalities online, corresponding to the level 3 or the most advanced user profile described in Section 5. In other words, when logging in, a PangeaMT SaaS Power user can see **Memories, Domains & Engines** menu completely activated and enjoy all powerful functionalities therein.

The PangeaMT platform managed by the PangeaMT team to cater for all levels of PangeaMT SaaS users, including Saas Power ones, works in dedicated physical and cloud servers and makes use of a number of sophisticated technical features in the background to ensure optimum and secure performance of MT work processes, such as heavy machine translating and engine training, being requested simultaneously by a significant number of users scattered worldwide. PangeaMT SaaS Power users may rest assured that their MT jobs, their TM assets associated to domains of their choice and their custom engines, which they can self update, are available to them round the clock and, most importantly, exclusively.

6. Conclusion

While the PangeaMT technology was borne out of Pangeanic's translation automation needs and expectations as an LSP, we soon decided to market it as custom, fully-tailored in-domain and client-specific engines. PangeaMT technology has evolved to go far beyond the custom engine creation and output-only offering portrayed by our competitors. We wished to set apart from other MT providers that impose word-based or user number limitations to their clients in the deployment of those engines. But was that the only means to empower our users? How about letting them schedule an automatic engine training without our intervention once their engine was first trained by us? Since 2011, users of a self-hosted PangeaMT DIY solution make the most out of their domain- or client-specific bi-texts by mixing and experimenting with these data sets for unlimited in-

domain or mixed domain SMT engines (e.g. safety/insurance and automotive, software and life sciences, etc.) and automatic engine retraining or updating.

The new PangeaMT platform presented here, released in 2012, has allowed us to become even more customer-focused and much more versatile in the array of MT-related services and solutions that we can now provide. Thanks to this year's revamping of the 2011 PangeaMT DIY concept in the form of an integrative, user profile-driven and multi-function platform, which can be installed at the client's end or accessible remotely on the Web, MT processes that were previously in the exclusive hands of the MT provider can now be elicited by the user when need be. This is so far the maximum exponent of our mission, i.e. to democratize MT – lowering or even removing access barriers to MT.

The PangeaMT team has strived to ensure that an organization that gets interested in MT but is a newcomer to the field can adopt the technology in an intuitive fashion, yet powerfully and independently from us as much as possible – and when that makes sense, that is, once they get acquainted with essential preparatory steps that we do gladly take care of, such as Data Consultancy, team-wide and management demonstrations and custom development project piloting.

Acknowledgements

The authors wish to thank their ever-growing worldwide and heterogeneous client-base for their invaluable feedback, particularly w.r.t. their different DIY platform and API related needs. The PangeaMT team would like to express their gratitude for the awarded European Union funds under the FEDER program managed by Valencia's local government IMPIVA in order to advance our Statistical Machine Translation technologies (award numbers: IMIDTA/2010/1016, IMIDTA/2011/777 and IMEXPB/2011/1). The collaborating UPV team feels indebted to Spain's Ministry of Science and Innovation (MICINN) and their own University's Research Council for awarding them the grants TIN2009-14511 and UPV/2009/2851 respectively.

Appendix: Last Word on Technical Matters

The PangeaMT Full Power platform is also functionally linked to the PangeaMT API, which allows for easy integration of a PangeaMT customization in computer-assisted localization and multilingual content consumption workflows and applications. Depending on whether you are the localization department of a corporation, an LSP or an organization in need of MT as an enabling technology to accomplish other multilingual technology tasks, you will need direct access to the Platform to query and retrain engines yourself or your own technology applications will be calling in your custom PangeaMT engines via API to get content translated automatically.

For those interested in knowing what is in a PangeaMT solution from inside out, particularly of interest to those clients in need of a self-hosted Full Power solution, PangeaMT works through the virtualization software VirtualBox Virtual Machine (VM). These prospects should get in touch for the latest information on the VM version in use, as well as the recommended technical specifications of the machine where the VM will be run, at the time of commissioning the solution. User's system requirements, such as compatible browsers, recommended resolutions and the like, will be of interest to both self-hosted and SaaS customers, and up-to-date information can also be provided at any time. While SaaS users do not have to bother about hosting technicalities, a natural concern that may arise in them is that of security, at the level of both data and engine handling. Having partnered with European secure hosting leaders, such as the prized company Strato, PangeaMT can ensure a high level of security at all times.

Should you require further details on technical matters, please do not hesitate to contact Alexandre Helle by e-mail. For a cost-free consultation on commercial aspects of a real-world PangeaMT customization as well as cross-company technology collaborations and integrations, kindly get in touch with either Elia Yuste or Manuel Herranz.

References

Papineni, K. et al. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the 20th ACL*. Association for Computational Linguistics. Issue: July, pp. 311-318.

Simpkins, A. (2012) Do-It-Yourself MT. In *Multilingual*. Multilingual Computing Inc. Issue: July/August 2012, # 129 Vol. 23 Issue 5, pp. 44-44. www.multilingual.com/articleDetail.php?id=1946

Yuste, E. et al. (2011) Going Hybrid: Pangeanic's and Toshiba's First Steps Toward ENJP MT Hybridization. In AAMT Journal. Issue 50: December 2011, pages: 33-39. ISSN 1883-1818.

Yuste, E. et al. (2010) PangeaMT – putting standards to work... well. In *Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas – AMTA 2010*, Denver. Available at: amta2010.amtaweb.org/AMTA/papers/4-04-HerranzYusteEtal.pdf

総会講演

国立国会図書館の電子図書館

長尾 真

1　国立国会図書館

国立国会図書館は立法府、即ち国会に属しており、東京本館、関西館、上野の国際子ども図書館の３つからなっている。現在の所蔵資料は、図書約970万冊、雑誌960万冊、新聞460万点、音盤、ＤＶＤ、その他を入れて合計で約3700万点の資料を持っている日本最大の図書館である。18歳以上の人は誰でも利用することができる。世界的には5番目前後の規模の図書館である。

2　国立国会図書館における電子図書館

国立国会図書館は国民の税金でまかなわれているから、そのサービスは日本中どこにいる人にも平等に行われることが理想であるが、現実には東京近辺の人達は気軽に来て使えるが、遠隔地の人達はほとんど利用できない状況である。これを理想に近づけるためには、所蔵資料を電子化してネットワークを通じて、全国に配信することである。つまり電子図書館を作ることが課題となる。しかし、ここには幾つかの難しい問題が存在する。資料のディジタル化には膨大なお金を必要とするほかに、現在の著作権法では著作権者の許諾なく、ディジタル化することができない。またディジタル資料のネット上の配信も著作権者の許可を必要とする。著作権が切れた明治時代の資料は自由にディジタル化し配信することができるが、その量は30万冊程度で、しかも古いものばかりなので利用者の数は自ずと限られることになる。

そこで著作権法を改正し、国立国会図書館に限っては著作権者の許諾なく、いかなる資料も自由にディジタル化できるようにした。ただ、その利用は国立国会図書館に限られるという制約が課されている。資料のディジタル化は2002年から少しずつ行ってきたが、最近127億円の特別予算を獲得してディジタル化を急速に進めた。その結果、図書については明治から1968年までの全てのもの、雑誌については1万2千種の雑誌の創刊号から2000年までのもの、古典籍7万冊、博士論文14万冊など、合計で現在210万冊の資料がディジタルの形になっている。

このような膨大な予算を使ってディジタル化した資料が国立国会図書館に来ないと見られないというのは残念だということで、文化庁の著作権分科会で議論が行われ、これらの資料で簡単に入手できないものについては、公共図書館、大学図書館等にネットワークを通じて配信してよいという合意がなされ、近く著作権法がそのように改正されることになっている。このように全国の公共図書館等への配信が実現することによって、日本中どこにいても最寄りの図書館に行けば、電子的に国立国会図書館の資料が見られることになる。このようなシステムは世界で初めてのもので、そういう意味で日本の電子図書館システムは世界をリードしていると言える。

近ごろは貴重な情報がウェブサイト上に存在することが多くなってきたので、国立国会図書館法（と著作権法）を改正し、国，地方公共団体、国公立大学、独立行政法人などのウェブサイトを許諾なく収集できるようにした。３．１１の大震災直後から東北地方の自治体などのウェブサイトを中心に毎日毎週収集を行ってきたが、これらは３．１１以後、各自治体等がどのような行動を取ってきたかと

いう貴重な情報を提供してくれている。

3　電子資料の検索利用

図書館の利用のために、これまでは書誌情報（OPACという）の検索が行われてきた。しかし、OPACだけでは資料の内容を推定することが非常に難しいので種々の工夫が行われてきている。国立国会図書館においては約1600万点の検索ができるが、そのほかに公共図書館や大学図書館、専門図書館、国立公文書館、国立美術館、その他のOPACまでを横断検索して、要求された資料が国立国会図書館のほかに検索者の近くの公共図書館にあれば、そのことを知らせるといった便利な機能も備えている。検索語のあいまいさを許すあいまい検索や、検索した書物についての新聞の書評へのリンク、またその書物を買いたいときはオンライン書店にリンクしてゆけるといった多くの機能を持たせている。

調べ方案内という機能によって、あるテーマについての種々の資料を参照することができる。たとえば“即席めん”について調べたいという場合は、その統計データの所在、その製造業のリスト、またこの業界事情といったことを調べる資料名を教えてくれたりする。雑誌については雑誌名や巻号などのほかに、掲載されている各論文の表題、著者などもリストアップされているので、これらを直接検索対象とすることができる。この雑誌記事検索は非常に便利である。

4　電子書籍の流通基盤

グーグルは既に1600万冊の図書をディジタル化し、世界最大の電子図書館を誇っている。アマゾンも電子書籍の販売に出て来つつある。これからの電子書籍の流通基盤をどのような形にしてゆくかが大きな課題として浮かび上がってきた。そこで筆者は2008年に出版社にとっても国民にとってもメリットのある電子書籍の流通基盤のモデルを提唱した。長尾モデルと呼ばれているものである。これは出版社が電子書籍を国立国会図書館のデータベースに送り込み、そこから買いたい人に販売するという方式である。また購入はしないが借りてちょっと見てみたいという人は比較的安いお金を払うことによって、その人の端末から一定期間（たとえば1週間）その本を読むことができるようにしようとするものである。読者が払うこれらのお金はすべて出版社、著作権者にわたるようにする。こうすることによって全ての電子出版物は国立国会図書館で永久保存され、出版社は自分でサーバを持たずに読者に出版物を売ることができるわけである。紙の本の場合は少なくとも何百部か、何千部か売れる見込みがなければ出版できないが、電子書籍の場合は、国立国会図書館のサーバに無料で預けておけるので、売り上げ部数に応じた収入が確実に入り、印刷、製本、流通等にかかる費用は一切必要ないわけで、大変有利な流通販売のシステムを構成できるのである。いわば国立国会図書館を中心としたクラウドシステムによる電子書籍の流通システムが実現される可能性があるだろう。

5　これからの電子書籍

読書はやはり紙の本でないとだめだという人は多い。しかしそのように言う人達は紙の本を電子形態で読むのを嫌がって言っているのであって、電子書籍の持つ多様な可能性に気付いていないからであろう。電子書籍の場合には、たとえば英語の発音を聞くことができるし、写真のかわりに動画像（すなわち映像）も見ることができる。すなわち電子書籍はマルチメディアの著作物なのである。これによって表現できることは紙の本のような2次元の世界から、音や映像などを含んだ4次元の世界に世界が拡大されるのである。

さらに読書端末から読者が書物に対して働きかけをすることができる。文字の拡大縮小は当然であるが、問いを発することによって、読書端末の方から答えを与えてくれる（富士山の高さは？と聞くと

3776 メートルと答えるなど）し、友達同士で１つの作品を読みながら意見交換をしたり、ネットを通じて著者と読者が対話をするといったことも可能になる。このように電子書籍は全く新しい次元の著作物と位置付けられるわけである。

6　理想の電子図書館

ディジタル化された書物は目次などにしたがって構造化することができる。そこで１つの書物のある部分が他の書物のある部分と密接な関係を持つといったことを、自然言語処理技術によって発見することができるだろう。こういったリンクは種々の因果関係について作ることができるから、電子図書館の本は単純に並べられて記憶されているのではなく、それらの本のいろいろな部分がいろいろな因果関係で他の本のある部分とリンクされるといったネットワーク構造が作り出されることになるだろう。これを究極のところまで進めれば、人間の頭脳における記憶の構造に近いものになってゆくだろう。そこでの検索は本を単位に取り出してくるだけではなく、必要とする本の部分が取り出され、またそれに関連する情報が芋づる式に取り出されることになる。これは情報検索でなく、事実検索（fact retrieval）の世界である。つまり人間頭脳内の知識の構造に近いものが電子図書館で実現されることになると考えられる。そこまで電子図書館を持って行くためには非常に高度の自然言語処理技術が開発され、巨大データに対して適用される必要があるわけで、電子図書館は自然言語処理の対象として非常に魅力のあるものと考えられる。

総会講演

展開が期待されるメディア社会におけるMT

東京工科大学メディア学部
飯田　仁

1. はじめに

2011 年 9 月に開催された MT-Summit XIII において，筆者が七人目の IAMT Award of Honor を受賞した．同賞は MT-Summit 1997 において初めて設置され，初代受賞者は AAMT 初代会長の長尾真先生である．1997 年の大会はコンピュータ登場 50 周年を祝う行事の流れの中で開催された．高齢となった MT-Pioneer たちが参加し，当時を語るパネルに登壇した．

本稿では，その黎明期からこれまでの MT の発展を概観して，これからの MT が進む方向について議論したい．とくに，2010 年代の社会活動は PC を主体とするインターネットからモバイル端末を使ったソーシャルネットワークへと変容している．国レベルでの多言語発信サイトもその価値を高めていくであろうが，モバイル／ソーシャル／リアルタイムという主たる概念を軸に 21 世紀が動き出しているとみる．技術面の発展に向けて，大規模コーパスの構築や統計的な知識モデルの構築などをそれらの軸空間の中で捉えてみて，世界規模のインタラクティブ性の実現が期待される MT について考える一助になればと思う．

2. MT-Summit 1997 に登壇したパイオニアたち

1946 年に N. Wiener と Warren Weaver とが，そして Weaver と Andrew D. Booth とが話す機会をもち，翌 47 年に Weaver が機械翻訳の取り組みを提言し，51 年に MIT にいた Bar-Hillel が機械翻訳の可能性を示唆した．コンピュータの登場と共に機械翻訳の実現が協議されたその時期に登場する D. Booth は 97 年の大会に臨むことはできなかったが，大会予稿集にメッセージを寄せている．そこでは，IBM のコンピュータの開発機を順次利用して，54 年から Georgetown 大との機械翻訳研究が始まった．その後，64 年までの研究費総額が 2,000 万ドルといわれている．この間，米国の外では，57 年の Sputnik shock が起き，同年には Chomsky が “Syntactic Structures”を出版した．また，Kennedy-Khrushchev 間の一時の米ソ雪解け外交があり，ソ連からのユダヤ人移民を米国が受け入れている．このことは 66 年の ALPAC report に遠からず影響を及ぼしているように思える．そのような話を MT-Summit 1997 のバンケットのお楽しみ抽選会の副賞として筆者が射止めた「パイオニアたちと懇談する時間」の 1 時間内で聞くことができた．同大会予稿集にメッセージを記載した主なパイオニアを表 1 にまとめる．表中の氏名に下線があるパイオニアが大会に参加し，直接話を聞くことができた先達である．

表 1 ：MT-Summit 1997 で紹介されたパイオニア

開始年	パイオニア氏名	所属，活動内容など
1946	Andrew D. Booth	Weaver と電子計算＆MT
1953	Victor H. Yngve	MIT, 独英 MT, '65-Chicago
1954	Michael Zarechnak, std. of R. Jacobson	IBM & Georgetown Project 参画
1954	Christine A. Montgomery	Georgetown 露英のデモ
1956	Winfred P. Lehmann	Georgetown とテキサスでの独英
1958	Peter Toma	Georgetown, 1961- 87 SYSTRAN(1969)

60 年代に入り，ソ連からのユダヤ系移民たちが露英翻訳で活躍したことが機械翻訳に比べ人手翻訳コストが廉価だとする ALPAC の報告と無縁とは思えない．結果，米国において，独英翻訳や SYSTRAN などの研究開発が継続された．

その後，ほぼ 20 年の歳月を経て，1987 年に第 1 回の MT-Summit が箱根で開催された．

3. MT が活きる社会環境の変化

機械翻訳が自動であれ，支援であれ社会が求める場面は多彩であり，テキストに限らず音声対話も対象となり，多言語への対応が期待されてきた．新聞，マニュアル・説明文書類，報告書類，グローバル社内文書，特許文書，窓口サービス案内，地域観光案内などはいずれも人手翻訳と自動翻訳とのコスト比較が算出可能な対象であろう．一方，21 世紀に入る頃からネット社会，ならびにメデイア社会の状況が大きく変化してきていると言える．とくに，Windows95 以降，その一端を書き連ねてみる．

1998	Google 誕生
1999	2 ちゃんねる開設
2001	Wikipedia 登場（英語版）
2004	GREE，mixi 運営開始（日本） FaceBook 運営開始（米国）
2005	YouTube 登場
2006	Twitter 開始（米国），ニコニコ動画登場
2008	iPhone 3G
2010	スマート TV
2012	new-iPad ，FaceBook ユーザ 8 億人突破，上場

以上のような環境の変化から人々の活動がソーシャルネットワーキングサービス SNS の上で進んでいる傾向が顕著になっていると言える．さらに，日本から世界に独自に発信し普及することを目指しているインターネット上のサービスとして，つぎのようなコンテンツやサービスが知られている．

- 歌声合成ソフト：「初音ミク」（英語版に展開），中国展開（ヤマハ）
- 絵文字：装飾メール素材のネット上の提供
- ソーシャルゲーム：ディー・エヌ・エーやグリーの中国・米国展開
- 電子書籍：コミック誌電子版の米国・カナダ販売
- 動画：ドワンゴ「ニコニコ動画」の英語つぶやきの米国配信
- 漫画などイラスト：漫画イラスト投稿サイトの 7 言語提供

これらのコンテンツやサービスの楽しみ方がグローバルに広がりを見せていることは，モノや地域に固定されることを超えた，つまり書物やテレビを介した社会との関わりでない，ネットを介した人々の繋がりや社会との繋がりが PC をも越えたモバイルツールの下で実現されつつあることを示す．モバイルに特化した SNS として，Instagram や foresquare などの利用者が増えている現状を捉えておくことは重要である．

4. これからの MT のあり方を考える

社会活動の形が変化していくのに伴って MT のあり方，役割を考えざるを得ない．十分な議論を要する必要があるのだが，ここでは，つぎの 4 つの点に言及しておきたい．つまり，これから発展が期待されるメデイア社会において，人間の行動と社会活動に MT が如何に支援し得るかを押さえておく．

① 自己完結型行動・指示： 必要性薄
スマホ音声入力など

② 世界規模での発信と偏在する情報解析： 多言語 MT
Web-論壇， 報道メディア論調分析など

③ 世界市場の消費活動支援： 特定同時 MT
クレーム処理など

④ ソーシャルメディアサービス： サイト内同時 MT
コミュニティ内の言語バリア解消など

①においては，自身の行動様式などのモデル化をすること，つまり自身の行動のパラメータ化ができれば，統計的なモデル化により対応可能であろう．③や④に対しては，コミュニティ内の行動様式やインタラクティブな行動様式のモデル化が必要であり，コミュニティ固有の特性を捉える社会学的，社会行動学的知見を取り込んだモデル化の取り組みが重要となる．そして，②については，コミュニティ内の行動様式を世界規模のグローバルな視点で捉えるモデル化を前提にすることが欠かせない．そして，これらのモデル化における情報の粒度との関係を捉えることが活きた応用，サービスを具現化する要点となると考える．これまで MT の研究とは独立して追究されている詳細な会話分析やマルチモダリティの分析研究の成果を十分に取り入れた統計的なモデル化が必要と考える．

5. おわりに

21 世紀の社会活動の 3 種の特徴を軸とする空間の下，活きた MT を追究していきたい．

機械翻訳ソフトウェア一覧

このページに掲載しているソフトウェアはAAMT独自で調査した、日本国内で販売しているあるいはアジア言語を対象としている翻訳ソフトです。辞書引きのみのツール、オプション辞書は原則として掲載しておりません。また、「+OCR」「＋辞書」という製品も掲載しておりません。
ソフト名称は、原則としてプラットフォーム/言語対/バージョンを省略しています。(例:「AAMT/ej for windows V2.0」という製品の場合→「AAMT」と記載)
英日/日英以外の言語については、必ずしも翻訳方向を示していません。
URLは、企業のトップページあるいは代表製品のURLの場合があります。
本ページに記載のソフトの使用、本ページに記載のURLにアクセスしたことに起因するすべての損害をAAMTは補償いたしません。
本リストに掲載されていないソフト、URLの変更などご存じでしたら、ご連絡いただければ幸いです。

順不同

2012年7月6日現在

会社名(略称)		対象言語			OS		
	ソフトウェア名	英日	日英	その他	W:Windows M:Mac OS		
東芝ソリューション (http://hon-yaku.toshiba-sol.co.jp/)							
	The翻訳プロフェッショナル	EJ	JE		W		
	The翻訳エンタープライズ	EJ	JE	中	W		
日本電気 (http://www.nec.co.jp/middle/meshplus/)							
	CROSSROAD for Enterprise	EJ	JE	中韓	W		
	CROSSROAD	EJ	JE		W		
沖電気工業 (http://www.yakushite.net/)							
	訳してねっと	EJ	JE	中日			
富士通 (http://software.fujitsu.com/jp/atlas/index.html/)							
	ATLAS	EJ	JE		W		
ロゴヴィスタ (http://www.logovista.co.jp/)							
	LogoVista Pro	EJ	JE		W		
	LogoVista メディカル	EJ	JE		W	M	
	コリャ英和！一発翻訳バイリンガル	EJ	JE		W	M	
	コリャ英和！一発翻訳マルチリンガル	EJ	JE	仏独伊葡西韓中露	W		
	コリャ英和！一発翻訳医歯薬	EJ	JE		W		
	コリャ英和！韓国語			韓日	W		
	コリャ英和！中国語			中日	W		
	コリャ英和！欧州語シリーズ	EJ	JE	仏独伊葡西露	W		
クロスランゲージ (http://www.crosslanguage.co.jp/)							
	PAT-Transer	EJ	JE	英日韓	W		
	Legal Transer	EJ	JE		W		
	翻訳スタジオLE	EJ	JE		W		
	明解翻訳	EJ	JE	仏独伊西葡中韓	W		
	MED-Transer	EJ	JE	EJK	W	M	
	Web-Transer	EJ	JE		W		Linux
	MAC-Transer	EJ	JE			M	
	多言語パック	EJ	JE	仏独伊西葡	W		
	翻訳ピカイチ	EJ	JE	独仏伊西葡中韓	W	M	
	翻訳ピカイチ メディカル	EJ	JE		W		
	オフィス翻訳ピカイチ	EJ	JE		W		
インパルス・ジャパン (http://www.impulse-jp.net)							
	MagicalGate	EJ	JE	英中日韓葡仏伊西独			

会社名(略称)		対象言語			OS		
	ソフトウェア名	英日	日英	その他	W:Windows M:Mac OS		
高電社 (http://www.kodensha.jp/)							
	翻訳J・E・T	EJ	JE		W		
	j・Seoul			日韓	W		
	J北京			日中	W		
	Jソウルパーソナル			日韓	W		
	J北京パーソナル			日中	W		
	ChineseWriter			日中	W		
	KoreanWriter			日韓	W		
	翻訳ウォーカー j・Seoul			日韓			Pocket PC
	翻訳ウォーカー j・北京			日中			Pocket PC
	翻訳ウォーカー JET	EJ	JE				Pocket PC
日本IBM (http://www-06.ibm.com/jp/software/internet/king/)							
	インターネット 翻訳の王様	EJ	JE		W		Linux
	Websphere Translation Server	EJ	JE	中韓	W		AIX,UNIX,Linux
ジャストシステム (http://www.justsystems.com/jp/products/honyaku/)							
	翻訳ブレイン	EJ	JE		W		
山野敏夫氏 (http://www.tcct.zaq.ne.jp/yamano/index.html)							
	トラちゃん	EJ		エスペラント	W		
ワードバンク (http://www.ashiya.ne.jp/rosetta.html)							
	Rossetastone	EJ	JE				
テクノウエア (http://www.bekkoame.ne.jp/~twc/index.html)							
	PROjectMT			英露	W	M	
ソースネクスト (http://www.sourcenext.com/)							
	本格翻訳	EJ	JE		W		
	超速通訳 ツージル	EJ	JE	中	W		
デバイスネット (http://www.devicenet.co.jp/pro/tabiec.html)							
	たび通EC	EJ	JE		W		
テクノクラフト (http://www.technocraft.co.jp/)							
	ロボワード	EJ	JE		W		
イーフロンティア (http://www.e-frontier.co.jp/translate/pkg/)							
	同時通訳	EJ	JE	中韓日仏独伊葡西	W		
H. Takahashi氏 (http://www.vector.co.jp/soft/win95/edu/se364617.html)							
	Yamato英和.NET Lite	EJ			W		
ワックドットコム (http://www.wac-jp.com/products/hangryu/)							
	韓流インターネット			韓日	W		
アイデント (http://www.e-ident.net/htm_show.html)							
	韓国語 HTML Translator			日韓	W		
	韓国ネット旅の友			日韓	W		
カシオ (http://k-tai.casio.jp/au/)							
	カメラ撮影翻訳	EJ	JE				携帯電話
イチベル (http://www.ichibel.com/)							
	Jibbigo	EJ	JE				携帯電話
キングソフト (http://www.kingsoft.jp/dictionary/)							
	キングソフト辞書	EJ	JE	日中	W		
創新ソフト (http://cssoft.co.kr/jp/)							
	ezTRANS			日韓	W		
	ezTalky CE			日韓			PocketPC
	ezTrans2001 Office Add-in			日韓	W		Solaris, Linux
	ezTrans Server			日韓	W		Solaris, Linux
訳星 (http://www.sinicave.com/pd_transtar.cfm)							
	訳星個人版			英日中	W		
	訳星企業版			英日中	W		
	訳星中日版			日中	W		

会社名(略称)		対象言語			OS		
	ソフトウェア名	英日	日英	その他	W:Windows M:Mac OS		
華建集団 (http://www.hjtek.com/)							
	華建多語訳通			英中露日	W		
Kingsoft (http://jshop.javvin.com/language/translationexpress.php)							
	金山快訳			英中日	W		
欧泰科技 (http://www.otek.com.tw/jpnOtek/Transwhiz.aspx)							
	訳経Transwhiz			英中日	W		スマートフォン
台湾訳龍 (http://www.hostran.com.tw/newweb/pro10-sys.htm)							
	Internet Passport			英中日	W		
Mabnasoft (http://mabnasoft.com/english/parstrans/index.htm)							
	PTRAN			英ペルシア語	W		
バビロン (http://www.babylon.com/)							
	瞬訳名人バビロン	EJ	JE	英仏独伊葡西欄中	W		
Larry Smith (http://targumatik.tripod.com/)							
	Targunet			英ヘブライ	W		
	TarguLite			英ヘブライ	W		
	maXim			英ヘブライ	W		
	Targumatik			英ヘブライ	W		
ITC Inc. (http://www.itc.com.tr/engl/cev.html)							
	Cevirmen			英トルコ語	W	M	
Bilgi PARKI (http://www.sametran.com/)							
	sama tran			英トルコ語	W		
Sakhr Software Co. (http://www.sakhr.com/)							
	Sakhr Enterprise Translation system			英アラビア			
Trident Software (http://translate.ua/us/pragma-6x)							
	Pragma			英仏独露葡カザフ, ウクライナ,ラトビア	W		
PROMT (http://shop.promt.com/)							
	Promt			英仏独西伊露	W		
	Druz' ya Goo-ru			英独露	W		
Linguatec Language Technologies (http://www.linguatec.net/products/)							
	Personal Translator 14			英中仏独伊葡西	W		
	Linguatec Translator Corporate Solutions			英中仏独伊葡西	W		
Cimos (http://www.cimos.com)							
	An-Nakel Al-Arabi			英仏アラビア	W		
	MLTS			英仏アラビア	W		
SYSTRAN (http://www.systransoft.com/index.html)							
	SYSTRAN Professional			15言語	W		
	SYSTRAN Personal			15言語	W		
	SYSTRAN WebTranslator			15言語	W		
	SYSTRAN Office Translator			15言語	W		
ATA Software Technology (http://www.atasoft.com/)							
	golden Al-Wafi R Arabic Translator			英アラビア	W		
	Al-Wafi			英アラビア	W		
	Al-Mutarjim Al-Arabey			英アラビア	W		
Translution (http://www.translution.com/)							
	Translution	EJ	JE	11言語	W		
Translation Experts (http://www.tranexp.com/)							
	NeuroTran	EJ	JE	43言語	W	M	
	PocketTran	EJ	JE	43言語			Pocket PC
	PalmTran	EJ	JE	43言語			Palm OS
SDL (http://www.translationzone.com/jp/solutions/automated-translation/)							
	SDL Enterprise Translation Server			11言語			
	Language Weaver	EJ	JE	33言語			

会社名(略称)		対象言語			OS		
	ソフトウェア名	英 日	日 英	その他	W:Windows M:Mac OS		
Smart Communications, Inc., (http://www.smartny.com/translator.htm)							
	SMART Translator			英仏独希中葡西	W		UNIX
Language Engineering Company (http://www.lec.com/)							
	LEC TRANSLATE			21言語	W		
	Power Translator			21言語	W		
	LEC Passport			21言語	W		
	LEC Client-Server			21言語	W		
Worldlingo (http://www.worldlingo.com/en/products/)							
	Worldlingo Translator	EJ	JE	33言語	W		
Rocky Mountain Learning Systems (http://www.rmlearning.com/28906.htm)							
	Instant Immersion Translator Deluxe	EJ	JE	16言語	W	M	
Hebrew World (http://www.hebrewworld.com/targumatic.html)							
	Targumatik Pro			英ヘブライ	W	M	
AppTek (http://www.apptek.com/index.php/products/product-list)							
	TranSphere	EJ	JE	21言語	W		
	Ambassador	EJ	JE	10言語	W		
	WebTrans	EJ	JE	21言語	W		
Taragana (http://taragana.com/products/)							
	Translator Plugin			41言語	W		
VirtualWare Technologies (http://www.allvirtualware.com)							
	LogoMedia Translate			英仏独伊西露日中韓葡	W		
	PARS			英露ウクライナ	W		

委員会活動報告

UTX (universal terminology exchange) FAQ

http://www.aamt.info/english/utx/faq.htm
（日本語版：http://www.aamt.info/japanese/utx/faq.htm）

機械翻訳課題調査委員会ワーキンググループ3

Basic

Creating a glossary would require a lot of effort. I don't want to do it! I don't want to waste my time and money!

Actually, you are wasting your time by NOT making a glossary. Writing/translating documents without a glossary is quite time consuming and laborious. If you create a glossary, you can save your time. If you have reliable information in the form of glossary that other people can reference, you will see fewer mistakes. Everybody (including you) doesn't have to spend time in checking up terms that someone else already knows. You don't have to wonder which term is the best every single time you have multiple alternatives.

If you choose to create documents or develop software without a glossary, you are also wasting money. No matter how much money you spend implementing wonderful features into your application, your user will not notice them if your naming of the features are inappropriate and inconsistent. Using a consistent glossary can prevent this situation.

Isn't it hard to create a glossary?

That's exactly where UTX can help. UTX drastically simplifies the creation and maintenance of glossary by providing minimum, simple rules.

Are you getting any money for maintaining UTX?

No. The members of the UTX team (i.e. AAMT members) are dedicated volunteers. The activity of the UTX team is financed by AAMT.

How can I find out more about UTX?

The brochure, specification, and sample dictionaries are available at <http://www.aamt.info/english/utx/index.htm>.

I don't want to share any glossaries with others!

It's a pity, but you may still get benefits of using UTX by easily merging other UTX glossaries into your own.

Creating a UTX glossary

Do I have to include thousands of entries to create a useful glossary?

Absolutely not! The UTX team's research has shown that **a glossary containing as little as fifty entries** is useful to enhance the quality of machine translation for a 4000-word document. How a UTX glossary improve the overall efficiency of human translation is more difficult to measure, but the benefit of glossary will even extend to the improved readability and comprehension of the readers.

What kind of terms should we include in a UTX glossary?

A UTX glossary should contain technical terms within a specific domain. The majority of such terms are compound nouns. Please refer to the brochure and specification <http://www.aamt.info/english/utx/index.htm> for the details.

How can I edit a UTX glossary?

UTX can be edited with any spreadsheet applications (such as Microsoft Excel or LibreOffice) or text editors that can handle UTF-8 (such as "Notepad" included in Windows operating systems).

Can a UTX glossary include sentences?

It can, but sentences are better handled by translation memory formats, such as TMX. We recommend excluding sentences from a UTX glossary unless it's absolutely necessary. Generally speaking, you should avoid including an excessively long term in a UTX glossary. By keeping the length of terms to a certain length, columns should be readable.

Is a UTX glossary high-quality?

A UTX glossary should be high-quality, because its entries are hand-picked, and it should be inspected by a dictionary administrator. By contrast, automatically generated raw glossary data contain many inappropriate entries that degrade the quality of translation. This situation could be refer to as "**big data, big noise**." UTX's term status property allows a dictionary administrator to authorize or reject terms collected from various term contributors.

Can I use UTX to normalize terms?

Yes. A detailed instruction will be provided in the future.

Do I have to pay AAMT to create a UTX glossary?

No. AAMT doesn't charge you for the use of the UTX specification.

Can I change or sell existing UTX glossaries?

It depends on the license included in the header of the glossary. The UTX specification recommends indicating the license of a glossary. Creative Commons <http://creativecommons.org/> is a good idea. Of course, you can declare any license to your dictionary if it is legally reasonable. You can keep it for your own internal use. But UTX glossaries can be more useful and rich if you share them!

It's impossible to choose only one translation for one term! How can I follow the "one word, one meaning" principle?

If you find hard to follow this principle, you might have one or two of the following problems.

1. Assuming that you are the author of the document, you might be using multiple meanings for a particular term.

You should avoid using ambiguous terms in technical documentation. It is not a good idea to use multiple terms for a single meaning or a single term for multiple meanings. For example, avoid using "terms" to refer to an agreement, especially if your main topic is terminology. If you use potentially ambiguous terms, such terms must be

clearly defined and differentiated to show their proper uses.

2. You are mixing multiple domains into one glossary.

In principle, one domain requires one glossary. If a translation project deals with multiple domains, for example, medical devices, you may need to have glossaries for medicine, machine, medical devices per se, and perhaps more. You don't want to use a single glossary for the entire project, because it is not compartmented and hard to reuse.

If entries from multiple domains are included in one glossary without a good reason, the situation can be called **"domain contamination."** Different domain requires different terminology. For example, a file and a window have different meanings in carpentry and the ICT domain. If you maintain one glossary for one domain, one translation is for a source term.

For Translation Client/Language Service Provider

Why do we want to use UTX?

Have you felt frustration that you don't get technical terms translated correctly? Perhaps all you had to do was simply creating a glossary.

If you don't have any plan of using UTX for machine translation, then you will benefit from the simple terminological management of UTX. With a proper understanding, agreement, and arrangement, you might be able to collect contributions of new target term candidates from individual translators. Then you can create and use your own glossary.

If you are planning to use RBMT, then you can quickly build high-quality user dictionaries based on the UTX glossary.

We are translating books and games. How can we use UTX?

In books or game software, you will encounter tons of terms, perhaps many of them being proper nouns. They could be names of characters, skills, items, places etc. These are actually all **"technical terms."** Without a glossary, how would you properly keep track of thousands of terms over several months of translation? The readers of your book will be confused, and the user of your game will be angry when they see incoherent terms. Also your translation is likely to involve many translators and checkers. UTX is very useful to standardize the use of terms across translators and checkers, with or without terminological tools.

Can I propose a translation project that uses UTX?

Sure! Please let us know your ideas using the contact form <http://www.aamt.info/english/utx/index.htm#contact>.

I don't see why UTX could improve translation productivity.

That's perhaps because you are not reusing UTX glossaries. They are most useful when they are reused and/or shared among various user, tools, and environments.

Do I need to have a style guide to create a good UTX glossary?

Doing so is strongly recommended to maintain consistency. A number of well-established style guides are available for English and other languages. For Japanese, JTF Standard Style Guide <http://www.jtf.jp/jp/style_guide/styleguide_top.html> can be used.

Machine Translation

Why is UTX tab-separated format instead of XML?

UTX is designed to be simple. It is so simple that a UTX glossary is viable with only three mandatory columns (source and target term, and part of speech). They are manageable without using XML.

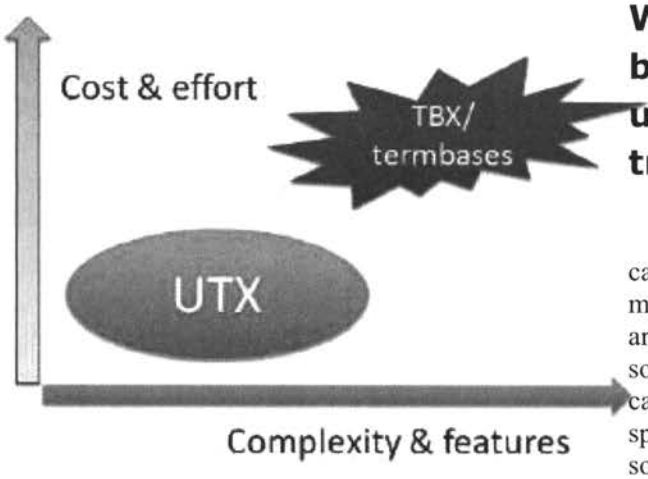
Why do we need a format if it is that simple?

Many online glossaries are published on the web, but many of them are very hard to use. They don't follow best practices of glossary. They often include similar entries without indicating priorities or clarifications of different usages. Their entries are not well-formed and they don't list their basic forms (singular or root form). However simple UTX looks, it can serve its purpose as a glossary by keeping to a certain specification.

Does UTX replace TBX, TBX-Basic, or any other existing glossary formats?

No. A UTX glossary can be created from scratch, as a collection of hand-picked technical terms by translators. It can be created with a very little effort (see the diagram below). It can serve as a basis for large-scale, complicated termbases for bigger translation projects. But it is quite useful as it is for small to medium-sized translation projects.

Position of UTX and TBX



What's wrong with TBX, TBX-Basic, or any other existing glossary formats?

There is nothing wrong about them. It's just that they are too complicated for a wider range of term contributors. Term contributors may or may not be familiar with XML or the details of various glossary formats. They can be professional translators who just know appropriate translations.

It would be nice to leverage such knowledge in the form of a usable glossary.

What is the difference between a system dictionary and a user dictionary (in translation software)?

(Rule-based) Translation software uses two types of dictionaries - system dictionaries and user dictionaries. A system dictionary is a collection of pre-defined terms that are fine-tuned to achieve the best translation results. A user dictionary is a collection of terms defined and added by the user to further increase the translation quality for a specific translation project. For this purpose, the entries of a user dictionary usually supersede those of a system dictionary. In general, a user dictionary should not include entries that are already included in a system dictionary. The user, however, can choose more suitable translations by adding such terms in a user dictionary and override the translations in the system dictionary.

What is the difference between a glossary and a user dictionary (of translation software)?

A glossary is a collection of technical terms that can be used by people or by software. A glossary may include definitions and descriptions, which are not used by translation software (translation software would need them in the form that they can understand). In contrast, a user dictionary is specifically created and used for translation software. One can convert a glossary into a user dictionary. At this point, the content of a glossary and a user dictionary is very similar. However, a user dictionary may have additional properties or entries that are not used by people. Generally, an extensive glossary can be a very good source for a high-quality user dictionary.

For developers (MT etc.)

Is the UTX specification established with RBMT (rule-based machine translation) in mind?

Yes. But UTX can be used with almost any translation/terminological tools.

Why did AAMT create the UTX format? What is the background?

Commercial translation software package like SYSTRAN is known worldwide, but you might not be familiar with translation software in Japan, where AAMT is based. The UTX specification is not limited to Japanese software or Japanese language, but a piece of historical background may be helpful to understand why UTX was established in Japan. In Japan, there are a number of commercial RBMT translation software packages. These high-end applications are shipped with 7-8 million basic/technological terms. They are highly sophisticated, and they have 30 or more options to control various aspects of translations (the high-end version of SYSTRAN has only 2 options for Japanese). As they can guess conjugations for user dictionaries, there is no need to feed detailed properties for each term entry.

Still, translation software need well-made glossaries to achieve good translation results. Large dictionaries could improve translation quality. They can, however, potentially degrade translation quality if the quality of the dictionaries is not adequately maintained. Our research proved that a small number of well-chosen terms in a UTX glossary significantly improve translation quality. This is the reason why we created a simple glossary format to reflect appropriate technical terms in translation.

We are using SMT. We don't need a glossary!

Perhaps you do. If you are using SMT and want to ensure the quality of translation, your project requires a separate process of terminological verification (which is integrated into the system if you are using RBMT instead). Even if you don't use a glossary when you translate, you will still need to use it for the purpose of quality assurance. Because you need a separate terminological verification process, you will need extra time and effort.

When converting to UTX, will it be a lossy conversion?

It depends. Although UTX can hold any amount of information by defining extra columns, doing so may not always be a good idea. If you need to maintain a number of extra properties, you may also need to consider the use of other XML-based formats. But we also need to realize that when we convert one format to another, only certain properties are essential.

Why does it not include lots of term properties?

Such properties contribute very little to improve accuracy/appropriateness of translation. Reducing complexity is more essential.

I would like to contribute glossaries/write a tool for conversion.

Thank you! Please let us know using the contact form <http://www.aamt.info/english/utx/index.htm#contact>.

Can I make suggestions to the UTX specification?

Sure! Please let us know your ideas using the contact form <http://www.aamt.info/english/utx/index.htm#contact>.

第 22 回通常総会および関連行事の報告

AAMT 事務局

当協会の第 22 回通常総会が 2012 年 6 月 19 日（火）13 時より・ホテルアジュール竹芝にて開催されました。総会後、各委員会からの報告会、講演会、そして第 7 回 AAMT 長尾賞授与式と受賞者による記念講演会が盛況のうちに行われました。

第 22 回通常総会

1. 開会の辞
2. 会長挨拶　豊橋技術科学大学・情報基盤センター教授・井佐原 均
3. ご来賓挨拶
4. 出席会員の確認
5. 議案

　　第 1 号議案　2011 年度事業報告（案）　第 2 号議案　2011 年度決算報告（案）
　　第 3 号議案　2012 年度事業計画（案）　第 4 号議案　2012 年度収支予算（案）
　　第 5 号議案　役員改選について（案）　その他・会員提案事項

6. 閉会の辞

報告会

開会挨拶　　会長・豊橋技術科学大学・情報基盤センター教授・井佐原 均

1. 機械翻訳課題調査委員会	委員長	長瀬 友樹	((株)富士通研究所)
2. AAMT/Japio 特許翻訳研究会	副委員長	江原 暉将	(山梨英和大学)
3. インターネットワーキンググループ	リーダー	富士 秀	((株)富士通研究所)
4. 編集委員会	委員長	宇津呂 武仁	(筑波大学)
5. MT サミット参加報告	監事	中岩 浩巳	(日本電信電話株式会社)

講演会

・国会図書館における電子図書館　　長尾 真先生
・展開が期待されるメディア社会におけるＭＴ　飯田 仁先生

第７回 AAMT 長尾賞授与式・記念講演会

受賞者：株式会社富士通研究所 機械翻訳システム研究開発グループ

▶　潮田 明　　富士 秀　　大倉 清司

受賞理由：

コーパスベースの言語資源構築技術の研究と開発、および機械翻訳システムの実用化技術に関する研究と開発における長年にわたる実績が高く評価できる。とくに、JST による「大規模多言語対応科学技術用語電子辞書の整備」におけるシソーラス作成による寄与、さらに特許対訳文の自動作成技術、翻訳効率が高い英語 Web ページ翻訳支援システムの開発、翻訳メモリの開発と商用機械翻訳システム ATLAS への搭載と実用化、総務省３ヶ年プロジェクト「国際情報通信ハブ形成のための高度 IT 共同実験」における多言語翻訳・翻訳支援システムの開発と実証実験などの多くの実績が認められる。

選考委員長：飯田 仁（東京工科大学）

選考委員　：横山 晶一（山形大学）　Virach Sornlertlamvanich（タイ NECTEC）
Key-Sun Choi(韓国 KAIST)　宇津呂 武仁(筑波大学)
隅田 英一郎(情報通信研究機構)

推薦者　：井佐原 均（豊橋技術科学大学）

同意人　：小谷 克則（関西外国語大学）

授賞式

懇親会

本会後の懇親会は、多数の参加者にお集りいただき、学識経験者、研究者、翻訳家等、幅広い分野の方々が活発な意見交換を行い有意義な交流の場となりました。ご参加誠に有難うございました。

協会活動報告

(2012 年 6 月～2012 年 7 月)

第 22 回通常総会

2012 年 6 月 19 日

第 1 号議案　2011 年度事業報告（案）　　第 2 号議案　2011 年度決算報告（案）
第 3 号議案　2012 年度事業計画（案）　　第 4 号議案　2012 年度収支予算（案）
第 5 号議案　役員改選について（案）
その他・会員提案事項

報告会

2012 年 6 月 19 日

①機械翻訳課題調査委員会　　②AAMT/Japio 特許翻訳研究会
③インターネット WG　　④編集委員会
⑤MT サミット参加報告

講演会

○講演：

・国会図書館における電子図書館　　長尾　真先生
・展開が期待されるメディア社会におけるＭＴ　　飯田　仁先生

第 7 回 AAMT 長尾賞受賞式・記念講演会

受賞者：株式会社富士通研究所 機械翻訳システム研究開発グループ

- 潮田　明
- 富士　秀
- 大倉　清司

受賞理由：

コーパスベースの言語資源構築技術の研究と開発、および機械翻訳システムの実用化技術に関する研究と開発における長年にわたる実績が高く評価できる。とくに、JST による「大規模多言語対応科学技術用語電子辞書の整備」におけるシソーラス作成による寄与、さらに特許対訳文の自動作成技術、翻訳効率が高い英語 Web ページ翻訳支援システムの開発、翻訳メモリの開発と商用機械翻訳システム ATLAS への搭載と実用化、総務省３ヶ年プロジェクト「国際情報通信ハブ形成のための高度 IT 共同実験」における多言語翻訳・翻訳支援システムの開発と実証実験などの多くの実績が認められる。

懇親会

2012 年 6 月 19 日　ホテルアジュール竹芝　12F　白鳳

決算理事会

2012年6月19日

第1号議案　2011年度事業報告（案）　　第2号議案　2011年度決算報告（案）

第3号議案　2012年度事業計画（案）　　第4号議案　2012年度収支予算（案）

その他・会員提案事項

機械翻訳課題調査委員会

2012年7月13日（2012年度　第3回）

① 前回委員会の議事録の確認

② 各WGの活動について（各WGに分かれて議論）

③ 活動内容の報告（各WGから）

④ 活動内容についての議論

⑤ まとめと次回委員会について

2012年8月27日（2012年度　第4回）

① 前回委員会の議事録の確認

② 各WGの活動について（各WGに分かれて議論）

③ 活動内容の報告（各WGから）

④ 活動内容についての議論

⑤ まとめと次回委員会について

AAMT/Japio特許翻訳研究会

2012年6月1日（金）（2012年度　第2回）

1. 前回議事録の確認
2. 特許文書の機械翻訳結果評価手法検討会関連（江原副委員長、Japio）
3. 第2回特許情報シンポジウム関連（横山副委員長、Japio）
4. 研究報告『自動評価手法IMPACTの実用化に向けて
   －評価時間の短縮のための改良－』（越前谷委員）
5. AAMT/Japio特許翻訳研究会ホームページ関連（事務局）
6. 次回の開催について

2011年7月6日（金）（2012年度　第3回）

1. 退任のご挨拶（JPO・森藤特許情報企画室室長）
2. 着任のご挨拶（Japio・松田事業管理室主幹／調査研究部部長）
3. 前回議事録の確認
4. 特許文書の機会翻訳結果評価方法検討会関連（江原副委員長、Japio）
5. 第2回特許情報シンポジウム関連（横山副委員長、Japio）
6. 研究報告『NTCIR-10　特許機械翻訳タスクでの評価の概要』（後藤委員）
7. 次回の開催について

編集後記

AAMT ジャーナル編集委員会委員長
筑波大学 システム情報系 知能機能工学域
宇津呂　武仁

AAMT ジャーナル 52 号をお送りします。

2012 年 1 月に、50 代の若さで急逝されました白井諭氏（NTT アドバンステクノロジ社）の追悼文を、中岩浩巳会長よりご寄稿頂きました。故白井氏は、NTT 研究所、ATR 音声翻訳研究所、NTT アドバンステクノロジ社におきまして、機械翻訳の研究開発において多大な貢献をされました。心からご冥福をお祈り致します。

今号に先立ちまして、2012 年 6 月に開催されました総会におきまして、長尾真先生ならびに飯田仁先生より貴重なご講演を賜りましたが、今号におきましては、両先生より、ご講演内容についての貴重なご寄稿を頂きました。

今後の巻頭言は、Translation Automation User Society（TAUS）会長の JAAP VAN DER MEER 氏より、ご寄稿を頂きました。また、TAUS における受賞者である Dieu Tran 氏（Cisco 社）、ならびに、Chris Wendt 氏（マイクロソフト社）より、受賞に関するご寄稿を頂きました。

一方、AAMT 内の活動報告として、機械翻訳課題調査委員会から、機械翻訳用ユーザー辞書共通フォーマット UTX（Universal Terminology Exchange）の英語版 FAQ を、インターネットワーキンググループから、機械翻訳ソフトウェア一覧を、それぞれご寄稿頂きました。

AAMT

Asia-Pacific Association for Machine Translation

AAMT 入会のご案内

AAMT は、機械翻訳の発展を目的として、機械翻訳の研究者、開発者、製造者、利用者が集まった任意の組織です。委員会による定期的な調査研究をはじめ、機関誌の発行、シンポジウム、セミナー等各イベントの開催など幅広く活動を行っています。
機械翻訳にご関心のあるすべての方にご入会をお勧めします。

＊＊AAMT 会員の特典＊＊

1.AAMT Journal の購読ができます。

会員には、機関誌である AAMT Journal （年 2～3 回発刊予定）が送付されます。購読料は年会費に含まれています。

2.機械翻訳関連の最新情報をメールでお届け

会員専用メーリングリストで、最新の機会翻訳関連の情報をお届けします。
MT 新製品、新サービスの紹介、国際会議、シンポジウムのお知らせ、WEB での MT 関連記事の紹介など盛りだくさんです。

3. AAMT が組織する委員会や調査活動に参加し、機械翻訳や翻訳に関心のある方との交流を深め、知見を広めることができます。

機械翻訳に関する言語資料の調査、広報、標準化活動に参加したり、AAMT Journal や会員専用メーリングリストで、自社製品、サービスの紹介を行うことができます。

4.関連機関の主催する国際会議に参加できます。

IAMT の主催で隔年開催される MT Summit をはじめ、AAMT、AMTA*、EAMT**の主催する会議やワークショップに参加できます。

AMTA* ：Association for Machine Translation in the Americas
EAMT** ：European Association for Machine Translation

年会費は以下の通りです。

法人会員：入会金　1 口　10,000 円　　年会費 1 口 50,000 円
個人会員：入会金　　1,000 円　　年会費　　5,000 円（学生は学生会費 1,000 円）

ご関心のある方は、事務局までお問い合わせください。

アジア太平洋機械翻訳協会（AAMT）
ホームページ：http://www.aamt.info
電子メール：aamt-info@aamt.info

AAMT

Asia-Pacific Association for Machine Translation

入会申込書

以下の通り、アジア太平洋機械翻訳協会の会員申し込みを致します。

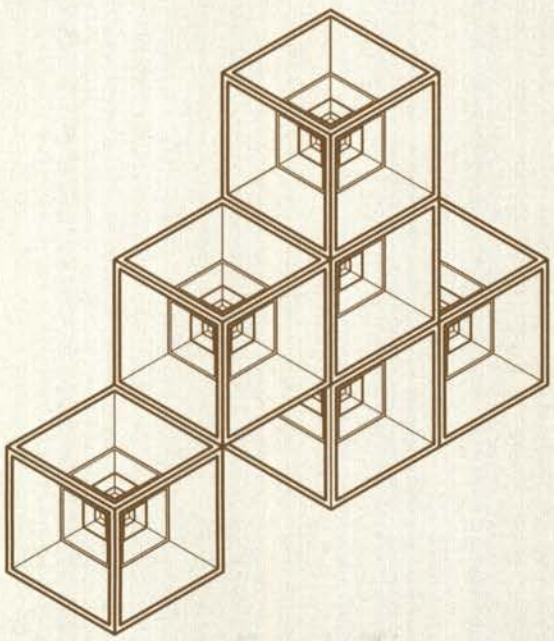
申込日　２０　　年　　月　　日

氏名（ローマ字）	
氏名（漢字）	
電話番号	
メールアドレス	
所属先	
所属先住所	〒
種別	ユーザ　　研究開発者　　その他
機械翻訳に関するお知らせメールの配信	希望する　　希望しない

コメント

Memo

AAMT



AAMTジャーナル No.52

発行：アジア太平洋機械翻訳協会（AAMT）
ホームページ：http://www.aamt.info
住所：〒441-8580　愛知県豊橋市天伯町雲雀ヶ丘1−1
豊橋技術科学大学　情報メディア基盤センター内
phone：0532-44-6620　fax：0532-44-6620
編集委員会：宇津呂 武仁　小谷 克則　大倉 清司
鈴木 博和　三浦 貢　村上 嘉陽
事　務　局：神崎 享子　服部 絹依
印　刷　所：株式会社ナビックス

Asia-Pacific Association for Machine Translation
c/o TOYOHASHI University of Technology Information and Media Center
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-City, Aichi, 441-8580 Japan
Phone:+81-532-44-6620　FAX:+81-532-44-6620
URL:http://www.aamt.info