Effectual MT within a Translation Workflow Panopticon

Sven C. Andrä
ONTRAM Inc.
10 South Third Street
San José, CA 95113, USA
sca@ontram.com

Jörg Schütz
bioloom group
Bahnhofstr. 12
D-66424 Homburg, Germany
joerg@bioloom-group.com

Abstract

In this presentation, we focus on integrating machine translation (MT) into an existing corporate localization and translation workflow. This MT extended workflow includes a customized post-editing sub-workflow together with crowdsourced, incentives based translation evaluation feedback routines that enable automated learning processes. The core of the implementation is a semantic repository that comprises the necessary information artifacts and links to language resources to organize, manage and monitor the different human and machine roles, tasks, and the entire lifecylce of the localization and translation supply chain(s).

1 Introduction

Today, huge amounts of continuous, multilingual data streams across time zones and cultural boundaries need translation processes and technologies that are able to be affected by their purposes and contexts, i.e. environment, origin, use case, etc., and their multi-dimensional qualities such as the diffusion distinction of being inbound or outbound. Even if the application scenarios have changed in these ways, any serious approach to mediate between various languages and cultures has to take into account the old and general wisdom of traditional localization attempts that:

i. Localization and transcultural adaptation are more than technical processes.
ii. While there is a strong technical component, successful localization in a global environment involves changing the way an organization does its business—see, for example, the recent discussions on Enterprise 2.0. As long as international markets are treated as a secondary concern and just a place to save on time and costs, globalization efforts will not be truly successful.
iii. In particular, localization is not a process that starts after a product has been envisioned and/or designed. If global concerns and plans for translingual after-market support are not made even before product development begins, costs will go up and quality problems will emerge.

This wisdom of mediation must be integrated into any modern interlinked environment to account for and to accomplish the requirements of today's globally networked industry and society.

ONTRAM, the Translation Management System (TMS) of Andrä AG, is a localization enabling platform and a globalization development framework that contributes to an effective and efficient solution to the above scenario. One outstanding innovation feature of ONTRAM is Translation Care, which is the fostering and the treatment of localization and translation activities and the prevention of globalization defects and immaturity. Translation Care is delivered by professionals in content production, terminology, translation, localization, internationalization, and allied globalization technologists from computer, information and media sciences. The ONTRAM Translation Care Repository (OTCR) pools, stores and indexes the complete language assets of the localization and translation production value chain with multiple information profiles and semantically enriched relationships.

In the following sections, we outline the integration of machine translation (MT) into an existing corporate translation workflow and how the OTCR contributes efficiently to an effectual overall translation automation deployment based on a strict semantic modeling approach of the processes, tasks and the various information artifacts as well as the involved actors and stakeholders.

2 Modeling Approach and Architecture

2.1 General Assumptions and Requirements

In our customer setting, we have the traditional multilingual product documentation workflow track which is augmented by a new, dynamic product support workflow in multiple languages with five so-called main languages within a customer support portal application that also facilitates multi-faceted navigation and natural language interaction.

In this new application scenario, customer inquires are seen as continuous data streams that need to be linked with, on the one hand, appropriate documentation, and on the other hand, already existing structured FAQ information—semantically tagged product information—across the different languages in order to provide an effective and beneficial product support as well as insights into possible product shortcomings. A strict language separation, i.e. the distinction between source and target language, would allow for entirely language-specific processes, but in our case the customer's language might not match the portal's UI language either because the language is not yet available or because the portal user switches the language or even switches between languages.

This opens up a web of information that in its basic configuration is similar to the ideas of the Semantic Web (SW) because the following basic assumptions apply to both:

- Extended AAA slogan (AAAA): "Anyone can say Anything about Any topic in Any language"[1].
- Open world in which we assume that there is always more information which could be known although our customer's world might be seen as closed world.
- Nonunique names which appreciates the reality that different users of the portal and on the Web might use different names to address the same entity (terminology mismatch).

These assumptions represent a fundamental departure from the classical assumptions of traditional information and knowledge systems architectures because the chosen mashup approach provides a highly dynamic environment in which information sharing can flourish and the network effect of knowledge synergy is possible. Similar to the SW, this style of information gathering, aggregation, syndication and sharing may generate a "chaotic" landscape full of possible confusion, disagreement and conflict if not properly routed and monitored which in our use case is possible because we have to deal with a near or quasi open world assumption (restricted by the actual product information).

Since human translations of the dynamic content would consume too many resources and could not be accomplished in near real-time, the decision was made to design an automated translation solution with a separate human guided correction and evaluation process to assure the basic quality requirement of "information completeness" for the translation results.

2.2 Selecting the MT Engine(s)

In 1988 at the TMI conference, Peter Brown and his colleagues of IBM introduced a statistical MT approach for French to English, and the audience was shocked because this approach neglected all the existing and emerging MT theories of many computational linguists (aka RBMT) at that time. Meanwhile SMT is the mainstream of MT research as well as of industrial scale employment, which also might be due to the fact that free translation service offerings of Google, Microsoft and others with SMT engines generated a wide spread momentum including acceptance and success.

Today we are also in the position that SMT can be trained on specific domain vocabularies and even on particular styles of writing which is mainly reflected through the employed language models and

[1] The AAAA slogan is an extension of the AAA slogan for the World Wide Web as introduced by Allemang and Hendler (2008): "Anyone can say Anything about Any topic."

translation models. These modeling processes benefit from different feedback channels which are essential for a successful training machinery, and therefore need continuous research into more elaborated and fine-grained modeling and feedback strategies.

Traditionally, RBMT can be improved through domain-specific lexical entities but the inherent rule shortcomings still remain in the engine. Feedback channels have therefore only a limited direct influence on the MT output because of various possible side effects. Nevertheless, the identified errors are more regular than SMT errors because of this internal rule regime. Repair operations are very time consuming and often additional side effects can occur very randomly. In addition, different qualities of the feedback data in terms of use and interpretation can occur.

Besides these criteria, the eventually selected MT system must also provide proper and seamless integratability into the employed Translation Management System to allow for effectively monitoring and controlling its processing.

To gain more insights into these kinds of processing environments, a first test instance had been set up with two SMT-based web translation offerings, one by Google and one by Microsoft. The initial decision for this proof-of-concept also included an additional separate test of the TAUS data sharing (TDA; TAUS Search) offerings for specific terminology mining tasks to keep the employed sets of language data up-to-date. Since any language system can be seen as a "living" system that as a pure data system requires constant updates because of new words and technical terms, new or additional meanings, new techniques and technologies, these aspects must be taken into account very seriously and with appropriate means to accomplish the open world assumption and the dynamic network effects.

2.3 Task-appropriate Post-Editing Process

Since today any MT output contains either minor or major defects due to the employed algorithms and data, a separate human post-editing task was added to the MT workflow. The post-editing (PE) task needs different capabilities to successfully cope with the MT output, and to also provide effective, usable feedback data for man and machine (O'Brien, 2006, and Schütz, 2008). A measuring system or metric had to be set up to carefully evaluate the productivity of the PE task in terms of performance and competence. The main aim is to receive informational correct and complete content that allows either fluent cognitive processing by humans or further semantic processing by machines of the entire information discourse of a given data stream. The throughput of the PE task has to be contrasted with an entire human translation approach for which we distinguish two incarnation strands: (1) rapid without separate proof reading (HT), and (2) fully fledged with a quality assurance step (HT++).

The first measure at hand is the classic "gold standard" of a time-quality ratio of these three working scenarios. If we assume the same translation quality result as defined in the subsequent validation scale then we obtain the following time and quality sequences:

- `t(MT+PE) < t(HT) < t(HT++)`
- `q(MT+PE) ≤ q(HT) ≤ q(HT++)`

which gives us enough evidence that we gained a productivity benefit (performance) without loosing much in quality (competence).

In addition, the post-editors ranked the MT output as well as the PE output in order to provide learning material for internal and external purposes —pre-processing training and possible SMT training. For this, each step in the task stream was evaluated on the basis of the following 4-level validation scale which represents our first best practice challenge:

i. language grammar is correct and the information is complete and correct—information content is fully processable
ii. language grammar is defective but the information is complete and correct—information content is "cognitively" processable
iii. language grammar is defective and some information is missing—information content is incomplete and therefore not directly processable
iv. language grammar is defective and information is wrong —information content is not processable

On level ii. and iii. we further distinguish between word or term mistakes including wrong verbs, and the possible change of meaning—for an in depth discussion see (Andrä & Schütz, 2010). The content of both levels and of level iv. is routed to the PE sub-workflow for correction and a final validation which also feeds back to the learning and training processes.

To successfully realize this process scenario and its work- and informationflows, it is necessary to take all stakeholders on board during the initial setup of the translation environment and the continuous maintenance and monitoring of the processes—these include content authors, terminologists, translators, proofreaders, project managers, copy-editors, developers, end-users, LSPs, technology suppliers, and many more—to define the process steps and tasks together with their genuine evaluation criteria along several dimensions which include language and informational quality scales in terms of performance and competence, costs, time, feedback capabilities and analytic capabilities.

From the initial test results, we derived guidelines as a basis for further best practice challenges for each role, step and task in terms of policy, governance and rules along the additional properties human, automation, productivity, fidelity, and so on. For the post-editor profile, we defined the following qualifications:

- Translator education
- Preferably bilingual source language competence
- Highly educated subject field background
- MT technology competence combined with an analytic MT mediation capability (MT affinity)
- Semantic modeling competence to support the validation of the repository content (see subsequent Section 3)

Our approach does not measure MT quality in general but measures how a set of particular tools and process steps matches set criteria in a given workflow scenario. It is therefore a multidimensional approach that is not biased in any direction. In addition, the approach also identifies different learning potentials for all involved parties and actors—humans and algorithms—and gives effective PE guidelines for the MT application scenario. This approach is entirely different to a blackbox evaluation approach which in essence automatically contrasts a machine translation result with a possible human translation, such as, for example, the major automated metrics BLEU, METEOR, TER and others including their derivatives.

2.4 Social Aspects

Through our customer's support portal social aspects enter directly the application environment, and therefore they have an important role which in the mashup scenario also extends towards the different actors in the entire life cycle of the application.

Social aspects cover a wide range of issues from interaction specific through information and knowledge sharing to ranking and assessment. Examples include:

- inform portal users about the technologies employed in the background, and ask for their immediate feedback with intuitive quick response facilities and fully fledged questionnaires
- all human actors in the internal processes are educated and trained with a particular emphasis on post-editors
- ranking tasks and their results on and the benefits for the algorithmic developments are visible to the contributors which in turn supports their motivation

The evaluation and feedback capabilities within the entire information life cycle and the derived value chain constitute an ecosystem that accelerates the network effect and fosters agile team playing and openness. The inclusion of the user community with their feedback and direct interaction enables a new generation of a localization and translation marketplace towards "personalized localization" and a successful transcultural language mediation.

3 Repository Realization

In this section, we give a brief overview of the actual realization and the implemented knowledge repository—the OTCR— with an event model that is beyond simple cause and effect as this is the case in traditional business process management repositories.

The key concept of our implementation is to leverage information assets that already have value for our customer such as existing schemas, controlled vocabularies, thesauri, FAQs, user inquiries and associated user and community feedbacks, etc., plus an information architecture and engineering model that on the one hand, applies standard engineering practices including the development of requirements definitions and test cases, and on the other hand, supports the modeling of semantic metadata that effectively guide and monitor the overall workflow and the different processes consisting of human and machine tasks and activities.

3.1 Repository Basic Design

The choreography language(s) with which we model all artifacts of the application including language resources and process and workflow realization of MT and PE that are executed within the Translation Management System ONTRAM are the standard SW formalisms of W3C. These models are the descriptive foundation of the ONTRAM Translation Care Repository, and the accessing and interpreting devices are standard, off-the-shelf inference engines.

Since the actual details of our modeling approach are beyond the scope of this presentation, we give only a sketchy overview of how knowledge gathering and extraction combined with the single inference operation are employed with the W3C standards:

- RDF provides a graph model and serves visualization purposes (subject-predicate-object triples)
- RDFS is used to express hierarchical relationships between artifacts and resources, e.g. “subClassOf”, as well as domain and range dependencies
- OWL provides further descriptive power to introduce, for example, sameness (“sameAs”) and more sophisticated relationships to model complete business processes
- In addition, the SKOS formalism might join in to relate traditional terminology approaches with the modeling of language resources, e.g. “broaderThan”

Besides the correction and streamlining of MT output, post-editors also validate the semantic annotations that have been generated automatically by the MT pre-processing steps, and ensure the quality of translingual annotations that are stored in the OTCR.

A further benefit we accomplish with this modeling decision is that it also eases and ensures the interoperability with other (future) third party applications because we solely employ the SW descriptive standards with the single but powerful operation inference. This approach is also future proof because recent developments of HTML5 and several microdata formats are in line with or can be combined with the chosen modeling languages. A further advantage of HTML5 is its ability to work offline due to its caching policy[2] which is an important feature for any web application.

It should also be mentioned that there are still problems with computational tractability and the overall complexity which are mainly due to the fact that this modeling approach is different from the traditional object-oriented modeling employed in various programming languages because of the intrinsic semantics of the SW languages—see for example the use of “domain” and “range” and its impact on inferences.

Regarding language related modeling problems, these concern, for example, languages which are semantically very different such as Chinese and English which arises from their completely unrelated origins in the Sino-Tibetan and European language families.

3.2 Workflow Semantic Triggers

The new MT connected workflow is completely modeled in the SW languages which was a quite forward task because in this workflow the dependencies are limited and the different process steps follow mainly a sequential execution line. We distinguish the following main process stacks that are defined by their relationship to the external MT service:

[2] To date, only Firefox 3.5+ and Safari 4+ fully support the HTML5 caching feature.

- unsupervised and semi-supervised MT followed by PE
- supervised MT with MT output annotation amendment followed by PE

Unsupervised MT means that the user generated content is routed to the MT service as is, i.e. without any labeling to support the MT processing. In the semi-supervised case, the content data stream is rudimentary labeled with metadata that is supported or interpreted by the MT service, such as for example domain information.

The supervised MT step consists of a preprocessing step in which customer- or product-specific terms are disguised according to a term encryption list, and a postprocessing step in which the terms are decrypted again and additional metainformation for the PE tasks is added. The term decryption mechanism does not change the MT result but labels the output accordingly.

At a later stage of the project, the preprocessing step can be enhanced with a rudimentary language proofing facility and entity recognition for further supporting the MT system (prepare for MT readiness) and metadata completion as well as metadata that give information on purpose, origin and context of the user generated data streams.

Some examples of these metadata types that are beyond pure cause and effect modeling patterns are :

- *Provenance* gives information about the source of a statement, e.g. "*Customer C reported that application A failed in environment E*." This type of pattern also includes information about the geographical location or place where a statement took its origin, e.g. "*The Singapore office reported that... .*"
- *Liklihood* expresses quantitative information such as probability, e.g. "*It is 80 % probable that application A fails in environment E.*"
- *Timeframe* gives time information about an artifact, e.g. "*Environment E was installed February 11 through March 6, 2010.*" This pattern also includes several other time related information elements such as timescale, timetable and time signature.
- *Context* expresses specific information about a project setting in which a statement holds, e.g. "*Customer C's IT set up environment E.*"

These types contribute to an effective guidance of the PE tasks and encourage the overall ranking of the MT results which ideally should also provide feedback to the MT services. However, how this feedback results in optimizing the MT services remains an open issue at this stage, and must be discussed in depth with the MT service providers.

3.3 Extensibility and Interoperability

New information elements can be added at any stage in the modeling process but this process needs always a careful evaluation because unforeseen inconsistencies or contradictions might occur in the model. Modeling defects are either caused by a misunderstanding of the underlying process states or the domain that has been modeled—remember the AAAA slogan. In most cases, the modeling tools for RDFS and OWL help with the identification of logical inconsistencies and other shortcomings.

It is planned to transform the existing technical documentation process landscape, i.e. the traditional localization and translation workflows, into semantic models, on the one hand to benefit from the new information architecture design, and on the other hand to allow for more effective interoperability because all business processes would be based on a semantic modeling approach and its associated tools.

4 Impacts and Results

4.1 Impact on Actors, Roles and Stakeholders

Given the nature of future communications and interactions between product manufactures and product users the web will further accelerate and continuously change information sharing, cooperation and collaboration with a strong emphasis on community use and organical growth. Semantic modeling supports these futures in three ways:

1. Framework for communication—man, machine, processes, workflows, etc.
2. Method for explaining conclusions

3. Structure for varying viewpoints

Because of these three aspects all involved parties of the proof-of-concept project shared a common understanding of the overall approach, and could successfully contribute with the knowledge of their respective domain—business analysts, process participants including translators, terminologists and post-editors, technical analysts, system administrators, and others.

4.2 Impact on Technologies and Vendors

The three ways of semantic modeling introduced in the previous section have also a growing impact on the developments of language and translation technologies and on how vendors offer their products and services in the future. In addition, the rapid evolution of the internet as a fully fledged ecosystem with its own rating and reputation capabilities and systems for anyone will also contribute to entirely new views on products and services in all daily application fields. In these scenarios openness and interoperability are key for market success.

4.3 Qualitative and Quantitative Results

Besides the obvious benefits of modeling business processes semantically, we finally summarize the achieved MT results, and the economic gains from different perspectives.

With an SMT employment, we encounter several traps that have to be taken into account on the informational level and in particular within the PE tasks. Besides grammatical mistakes that mainly concern wrong or non existing inflexions, we observed mistakes on the word level such as for example “switch on” vs. “switch off”, “boundedness” vs. “unboundedness”, etc. These mistakes are already well known when compared with Translation Memory traps which are quite similar. From this observation, we must derive a best practice to either label these situations for the PE step or for the training of both translation applications.

On the economic side, possible savings are in general:

- Time: the minimum being 20-30 % for MT+PE compared to HT and HT++, and the maximum being 50–60 % and even more
- Costs: up to 30 % and more depending also on the MT service agreement or license model. In our scenario this aspect was neglected because we employed the available free services. However, future fully fledged MT deployment must take this into account because any web service engagement has to validate trust and privacy issues that are associated with such services.

A PE profile is important and should be taken into account very seriously with particular aspects being community inspired crowdsourcing approaches and a proactive user / customer involvement which both provide feedback and ranking in several dimensions.

An often not discussed aspect concerns the staleness of data in all information pools and is related to an effective management of Translation Memory, MT, as well as OTCR content with efficient update routines.

Regarding research and development, our observation is that it is advisable to combine the different translation automation approaches with self-learning and self-sustaining system capabilities and to integrate them more closely with PE environments to better support post-editors.

Last but not least, why did we use the term “panopticon” for our architectural design? It was chosen to visualize the conceptual idea of transparency behind the overall design concept which allows an observer to observe all processes without the process executors being able to tell whether they are being monitored at a given point in time. Besides the term's traditional use for hospital and prison architecture designs, the term has also been used recently for many web user profiling and surveillance activities.

5 Conclusions and Perspectives

The ONTRAM Translation Care Repository is an innovation that helps to drive the entire localization and translation life cycle on the basis of semantic metadata and allows for a seamless integration of different translation automation technologies on the same description level. The approach is based on existing industry standards which are merged with the modeling standards of the W3C for the Semantic Web, and deploys already existing tools and engineering techniques from these fields.

Within a proof-of-concept project, the practical employability within a technical customer support portal has been successfully demonstrated. The implementation included the integration of machine translation, post-editing and feedback cycles for man and machine into existing workflows dealing with static technical information and dynamic, continuous data streams in multiple languages.

Since the proof-of-concept did not touch all the details that might occur in the wild of the application scenario, there remain many challenging potentials for future research and development work in the field of semantic metadata modeling together with translation automation workflows.

For our customer, the proof-of-concept opened new perspectives and trends which could be implemented on a broader cross-department scale. For them the lessons learned are in particular the power of an entirely community based approach, meaning to take all stakeholders on board, and the semantic modeling approach with the single and powerful operation of inference.

The lessons learned for us include the intrinsic complexity of semantic metadata modeling within this application area, in particular the what, when and why in terms of modeling patterns and anti-patterns, and the need to intensively evangelize a semantic technologies based approach throughout the industries.

References

Dean Allemang and Jim Hendler. 2008. Semantic Web for the Working Ontologist. Morgan Kaufman Publishers, Elsevier, Burlington, MA, USA.

Sven C.Andrä and Jörg Schütz. 2010. Employing Machine Translation in Glocalization Tasks – a use case study. Proceedings of the 14th Annual conference of the European Association for Machine Translation (EAMT), May 27-28, 2010, Saint-Raphaël, France.

Sven C. Andrä and Jörg Schütz. 2009. Cloud Computing in GILT Ecosystems and Evolution. ASLIB Conference Translating and the Computer 31, November 19-20, 2009, London, UK.

Sven C. Andrä and Jörg Schütz. 2009. MT Bazaar: translation ecosystems in the cloud. Proceedings of the Twelfth Machine Translation Summit, August 26-30, 2009, Ottawa, Ontario, Canada; pp. 395-402.

Peter Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, P. Roossin: A statistical approach to French/English translation. 1988. Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI), June 12-14, 1988, Carnegie Mellon University, Center for Machine Translation, Pittsburgh, Pennsylvania, USA.

Google Translate – http://translate.google.com

HTML5 W3C – http://www.w3.org/TR/html5

HTML5 WHATWG – http://www.whatwg.org

Philipp Koehn. 2010. Statistical Machine Translation. Cambridge University Press, Cambridge, UK.

Kathryn Blackmond Laskey. 2008. MEBN: A language for first-order Bayesian knowledge bases. Artificial Intelligence, Volume 172, Issues 2-3, February 2008, Pages 140-178.

Microdata – http://www.microformats.org/

Microsoft Translator – http://www.microsofttranslator.com

Sharon O'Brien. 2006. Machine-Translatability and Post-Editing Effort: An Empirical Study Using Translog and Choice Network Analysis. PhD Thesis, School of Applied Language and Intercultural Studies, Dublin City University, Dublin, Ireland.

Jörg Schütz. 2008. Artificial cognitive MT post-editing intelligence. AMTA-2008. MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas, Waikiki, Hawai'i, October 21-25, 2008; pp.448-453.

TAUS Search – http://www.tausdata.org/index.php/language-search-engine

TDA (TAUS Data Association) – http://www.translationautomation.com

W3C Semantic Web – http://www.w3.org/standards/semanticweb

W3C Webstorage – http://www.w3.org/TR/webstorage