Employing Machine Translation in Glocalization Tasks
– A Use Case Study –

Sven C. Andrä

Andrä AG

Torstraße 61

D-10119 Berlin

sca@andrae-ag.de

Jörg Schütz

bioloom group

Bahnhofstraße 12

D-66424 Homburg

joerg@bioloom.de

Abstract

Today, we and our customers are faced with a huge amount of continuous data streams in multiple languages and different forms and formats. Therefore, our business communications requirements and strategies demand for an effective employment of various language resources to economically and efficiently administer, master and monitor information processes and workflows across languages, cultures and time zones. We have thoroughly investigated into what language resources are mostly suited for our needs, and what are the important enablers in different translingual technical deployment scenarios that guarantee throughput, scalability, quality and successful operations and applications. Although Machine Translation (MT) is still a gadget because neither individual nor business users do share the usability and quality of MT as a real user experience, MT is an intrinsic part of our solution. With this paper we want to share and discuss our findings on MT with the community.

1.   Global Communications Landscape

In the last decade, private and business communications have changed dramatically with the Internet being the ultimate communications platform for everyone across time zones, languages and cultures. Translation and cultural adaptation play an ever increasing critical role in this global communications landscape and are no longer restricted to business and technical communication only. With the ubiquitous web access from different even mobile devices such as smartphones, the need for competent and effective language services increases exponentially, and in particular these services shall be highly configurable and available from everywhere, on demand, and preferable as a pay-per-use service offering.

Obviously, this new, highly proactive communications landscape with its associated demands for multiple language services cannot be handled, administered and controlled properly with traditional translation technology service set-ups because currently they are not flexible enough to account for the various communications requirements. The existing translation management systems are mainly designed to manage and monitor human-oriented tasks which in general are comprised of single, disruptive steps without direct interprocess relationship and interaction. Full automation and ambient adaptability are key to keep pace with the speed and variety of the multilingual transcultural demands, and the intrinsic characteristics of the processes from start to finish with persisted states. The term "glocalization" that is derived from the Japanese term "dochakuka" meaning "global localization" names the just described global communications landscape most appropriately.

In Section 2, we outline a possible technical solution that guided our language technology investigation with a focus on translation automation in particular machine translation (MT), and show what is needed to serve the demands of a global communications landscape. How this solution might be set into operation is discussed in Section 3, which also points to some serious shortcomings that exist with state-of-the-art technical and technological MT offerings with a focus on sharable language resources. In Section 4, we conclude and list the necessary steps in terms of short-, medium- and longterm investments, and how the MT

community – developers, providers, researchers and users – should contribute to these investments. Section 5 closes with additional perspectives for the decade and beyond.

2.   Technical Framework Solutions

A new breed of translation management framework is needed that also accounts for the effective modeling of the processes and workflows in addition to the efficient management of highly adaptable translation automation functions. The modeling and management aspects must be taken into account in an integrated way, and allow for effective and efficient automation and scalability from personal to business orchestration (local) and choreography (global).

In general, the core of such a framework is an event-driven workflow engine that offers capabilities to model, administer, monitor and test single processes and process stacks, and includes means to facilitate and continuously control human and machine feedback cycles which shall also serve as machine learning and training instances for amending and improving the services.

This engine can be based on an already existing translation management system workflow engine with some thorough redesign and extensions that account for the new needs and demands of a proactive, mobile, and multilingual information society and their specific language requirements. For example, appropriate APIs and SDKs[1] shall permit the integration of different third party offerings into a seamless process workflow, from modeling tasks down to the execution level, and its transparent management including the simplification of particular compulsory rules of regulation, governance and compliance.

For example, a possible deployment scenario might include content authoring support with language proofing and translatability monitoring as well as terminology services to ensure consistency and compliance with predefined corporate rules and conventions, or a post-editing guidance service in tandem with a machine translation deployment application.

As with other (business) processes, frameworks and tools that allow for an effective modeling and managing are needed to ensure an effective automation of various language and cultural adaptation services including their effectual qual-

ity monitoring on different levels and with multiple quality features which depend on the selected applications and their intrinsic processes.

In addition to the straight forward choreography of localization and translation services, the needed language resources of the involved different language technology services are no longer restricted to being a time consuming human effort but an automated continuous information extraction and discovery task.

The potential of this latter application also includes some of the biggest challenges and opportunities of the information age, namely connecting and integrating large sets of disparate data to create new sets for employment and analysis in a variety of applications. These challenges appear everywhere from large companies with several databases to data on the web for business and personal use. The problem of having information that is cleanly separated and largely inaccessible as been often referred to as the "information silo problem." In the microcosm of glocalization these challenges multiply with the number of languages and cultures we have to deal with. One solution is a better integration and a combined workflow management of the different services involved in the various glocalization tasks with built-in feedback capabilities. In addition to solving multilingual transcultural communication challenges, some of the problems with data integration and their process modeling and management are also dealt with and thereby contribute to a better targeted reuse and overall quality of multilingual data and information.

It also should be mentioned that only with such a technical framework solution for glocalization process modeling and management the TAUS forecast for 2010 (TAUS, 2010) gets its appropriate foundation to allow for

- A thousand MT systems will begin to bloom

- Sharing translation memories will take off in a big way

- Translation memory systems – as we know them – will cease to exist

Within the just started new decade of the 21st century, we will experience the evolution of language resources ecosystems with metabolic and regulatory pathways for which data models are necessary to effectively analyze and diagnose the

---

[1]API is the acronym of Application Programming Interface, and SDK stands for Software Development Kit.

ecosystem, and to provide appropriate tools and techniques for its effective operations and its overall control. Because MT is one of the most outstanding tools within these ecosystems, we will concentrate on how they fit with the described technical framework solution of process and workflow modeling, management and control.

3. Translation Automation

3.1 Use Case Scenario

In general, the employment of standard off-the-shelf MT as well as of situated and optimized MT mainly results in translation quality that is not appropriate in terms of vocabulary selection and grammatical choice in many applications.

For our information exchange requirements, we have evaluated translation results which we obtained through freely available Internet language resources (MT and TM) for several languages with a particular focus on German-to-English and German-to-French along these critical dimensions to get further insights about possible optimization and repair strategies.

This analysis of the MT quality shortcomings opened a relatively straightforward but computationally complex solution for repairing and enhancing the quality of the translation results through the employment of techniques from the field of computational intelligence. In Section 3.5, we outline our strategy and the chosen engineering approach. This approach makes use of sharable TM content and MT which we will also introduce briefly.

In addition, we have collected further requirements for the overall translingual environment which are essential for moving towards our vision of glocalization. These requirements concern the overall execution framework as well as the functions and capabilities of the employed automation components, and the mechanisms to ensure and assure the output quality. In the following subsections, we briefly describe the main building blocks of our use case scenario solution.

3.2 Process and Workflow Management

Traditionally, the management of the translation workflow is embedded in an relative intuitive system for assigning translation related tasks to the various people involved in the translation process and to monitor the overall flow execution.

With the demand to integrate more and more technical components and subsystems into this process flow and the need to evenly create process stacks of different kinds, the next generation of translation management frameworks must evolve to provide a complete process environment that is complex but readily understandable by all framework users, from the task analysts who create the initial drafts of the processes, to the technical developers responsible for implementing the technology that will perform these processes, and finally, to the people who are involved in some of the processes as well as to execute, manage and monitor these processes.

Thus, this process environment enables and creates a standardized pathway and a bridge for the gap between the process design and the actual process implementation by humans and by machine.

3.3 Translation Memory

For more than two decades the content of translation memory (TM) systems, which consists of aligned bilingual high quality translated segments, enables translators to faster and better build translations as well as to increase the overall translation consistency and quality. Furthermore, this technology enabled translation project managers to get to market faster and to control costs better.

Today, new approaches such as Advanced Leverage create new opportunities to effectively use TM content by employing more fine-grained segments together with linguistic and in particular semantic information.

In our use case scenario, TM content contributes to the overall quality of the results of the chosen engineering approach.

3.4 Machine Translation

MT, although more than four decades on the commercial market, has never reached the employment saturation of TM systems, and has not yet delivered the often promised translation throughput and quality. We distinguish mainly two strands of MT, rule-based MT (RBMT) and statistical MT (SMT). Recently, a third approach that combines the two basic strands has entered the MT stage.

The essentials of RBMT are thoroughly elaborated descriptions of language rules (syntax and to some extend semantics) and lexical data; and al-

gorithms that cover morphological and syntactical analysis and generation with a transfer step, the actual structural and lexical translation, between these execution branches. These rules foundered on the ambiguities of real human languages, and a lot of exception rules must be introduced which make any RBMT system very hard to maintain and to improve.

The essential steps of SMT to establish different statistical correlations and relationships in the translation process are load the statistics, translate the examples, evaluate the translations, tweak the system parameters, and repeat. The training materials consist of huge amounts of TM-like data for the statistical translation models as well as monolingual data for the language models of the source and target language.

To a certain degree, we might compare the statistical approach with the human brain because some of our "intelligent" language processes are apparently similar to a statistical inference engine. For example, our senses routinely make up for "noisy" data by interpolating and extrapolating whatever pixels or phonemes we can rely on.

Statistical analysis makes better sense of more data than strict rules do, and statistical rules produce more robust outputs in the sense of traceability. So any ultimate near human-quality translation engine must use certain statistics at its core, for example, to determine the actual context of a text to trigger lexical selection. In principle, the SMT approach converted text translation into an entire engineering problem, and employs a software architecture that allows iterative improvements. The built-in optimization architecture lets you swap out algorithms for better ones routinely, and the algorithms will change as performance improves.

In our use case scenario, we make use of all these MT system incarnations without any preference because we build our engineering solution on the actual translation results, and we do not care about how these results have been obtained internally. Nevertheless, the provided feedback cycles of our processes might have an influence on how the internal translation strategies of these MT systems will further evolve towards self-learning capabilities and ambient adaptability.

3.5   Translation Quality Metric

To measure MT quality, mainly the Bleu metric (Papineni et al., 2002) in several variations is em-ployed in different MT development scenarios to continuously improve the overall system quality. However, this metric is apparently not really helpful to measure translation quality in real life translation deployment environments because some MT systems are apparently just optimized for this metric with a certain reference corpus. The shortcomings of Bleu are also discussed thoroughly within the different MT research communities (see, for example, Callison-Burch et al., 2006).

Practically, a high Bleu score of an MT system means nothing in a particular context, especially in an industrial deployment scenario. In a competition, a win mostly proves that the winning system gamed the scoring system better than the others did. However, the worse the metric, the less likely the translations will make sense anyway.

What Bleu really measures is essentially word-by-word similarity: are the same words including sequences of words (the so-called n-grams) present in both documents, i.e. the MT output and the "gold-standard" human translation, somewhere? In obviously extreme cases, Bleu works pretty well; because it gives a low score if the documents are completely different, and a perfect score if they are identical. But in between, it can produce some very weird and bizarre results.

The most obvious problem is that paraphrases and synonyms cannot be taken into account by the measuring algorithms because there is no actual information that would trigger the identification of these relationships[2]. Therefore, to get any credit with, you need to produce the exact same words as the reference translation has. The complementary problem is that Bleu can give a high similarity score to nonsensical language which contains the right phrases in the wrong order.

For some purposes, perfect translation may not even be necessary, and for some applications, it might be enough to machine-translate most of the right words in mostly the right order, leaving to users the much harder task of extracting meaning from them.

In our use case scenario, a variation of the Bleu metric is used. Firstly, to use a currently adopted community standard for measuring MT output quality. Secondly, to guide the identification and selection processes with potential clue in-

---

[2] It should be noted that research is underway in this direction, for example, the METEOR approach of Banerjee and Lavie (2005).

formation for the scoring of the MT output. And thirdly, to accomplish the information gathering task for the feedback propagation to the MT system.

3.6 Best Practice Case Solution

To at least partly solve the terminology selection problem and the grammatical choice problem, we make use of several MT Internet services with a minimum of at least 3 services for each translation task. The different MT services are seen as independent translation ecosystems (Andrä and Schütz, 2009a) that contribute to an overall optimization problem that results in the final translation solutions.

In the ideal case, these solutions, either managed or unmanaged, should be send back to the services to accomplish a feedback propagation task that helps the service providers to improve their overall MT performance. In our view, these feedback cycles are an important step because they provide a mechanism that supports self-learning and self-adaptation capabilities of the MT systems which so far have been neglected by the system developers.

In the following, we do not discuss security and privacy aspects that might cause severe problems with the Internet translation services because it is neither transparent what data the providers store on their servers nor are the communication lines secure or the transmission is encrypted.

In our scenario, the actual translation process operates in five steps on an input I of a source language SL. In I all term occurrences are tagged during the content creation process:

- Call n MT Internet services, n > 2, with input I, and get back translations τ_1 to τ_n for the target language TL. Some few translation services already provide appropriate interfaces to automate the calling process, and all employed services can handle several input formats and they preserve the tagging in I.

- Check I against a (local) TM service, and fetch possible weighted matches tm_1 to tm_m

- Augment each τ_i with possible terminology information extracted from the tm_j results to establish a heuristics for possible term substitutions, and apply a TM matching algorithm on possible τ_i-tm_j pairs

- Validate each translation result t_i and tm_j including possible term substitutions and pairings against each other with a basic artificial immune system algorithm, and fetch a set of possible surviving translations. In this step, the defect type or the "infection degree" is determined.

- Optimize the set of surviving translations of the previous step by a swarm intelligence software routine. This step tries to cure the "infection" through eliminating the most severe defects.

We exemplify the essentials of these five steps with the following example sentences I_1, I_2 and I_3 in the German language[3].

- **I_1:** *Aufgrund seiner <term>Schwerhörigkeit</term> war es nicht möglich, sich mit dem <term>Patienten</term> zu verständigen.*

From six Internet services only one service delivered an appropriate English translation of the term "*Schwerhörigkeit*" in I_1, namely "*deafness*". All other delivered either the source term, or some bizarre translation such as "*hardness of hearing*" or "*tready bondage*". The term "*Patient*", however, was equally well translated by all services as "*patient*", which certainly is due to the term's general language use. The following sample sentences with a very dense terminology, nevertheless, were lost in (machine) translation:

- **I_2**: *Für <term>Patienten</term>, die neben ihrem <term>Husten</term> über ein <term>kratziges Gefühl</term> im <term>Hals</term> klagen, sind <term>Hustensaft</term> und <term>Lutschpastillen</term> geeignet.*

- **I_3**: *<term>Ambroxol</term>, <term>Lidocain</term> oder <term>Benzocain</term> wirken <term>lokalanästhetisch</term>, <term>Cetylpyridiniumchlorid</term> und <term>Dequaliniumchlorid</term> sowie <term>Chlorhexidin</term> und <term>Hexeti-*

---

[3] The full information content of the term-tag is not shown in the example sentences.

*din</term><term>desinfizierend
</term>.*

It should be noted that in many of these cases the statistical approaches among the MT Internet services delivered better term translations because their word selection process is apparently better guided by the term's contextual environment than the lexical selection process of the rule-based systems[4].

Step 4 and step 5 are further supported by a Web term crawling application that establishes an information shadow for each term occurrence so that further evidences for a specific term selection in the foreign language are gathered, and they account for the statistical extraction of a term's specific meaning. These meanings are stored in RDFS[5] triple stores within a domain classification system for future support of the content creation process, and steps 2 and 3 of the extended translation process.

Initial evaluations with the Bleu metric against terminological correct human translations show improvements of between 50 % to 100 %. However, it is too early to claim an overall success of our engineering approach, and more tests, algorithm tweaks and further evaluations are necessary. Nevertheless, the approach demonstrates the usefulness of a new breed of sharable language resources and open interfaces for their actual employment.

As a further optimization step we also envisage to employ human post-editing to optimally enhance the content of the TM systems with new translation pairs, and to effectively feed the MT services with quality-proven repair information in feedback routines if the translation service includes or offers such a feedback propagation option (Schütz, 2008).

Now, how does everything fit together? Firstly, with the terminological information the domain is identified. This domain information is essential to support the machine translation services in selecting the appropriate vocabulary base or even a separate terminological resource which might be offered locally within the company or by yet another cloud service. Obviously, this resolves most of the identified terminology problems in the MT environment. Ideally, this domain-specific information should be fed back always to the MT services so that future domain information can effectively trigger an appropriate vocabulary selection. The feedback information must be directed towards the actual shortcoming in order to serve as a competence and performance enhancer of the employed MT system, and this means the translation service has to provide an appropriate feedback interface.

Secondly, the grammatical choice problem of the MT results is partly solved by observing and analyzing the contexts within the different MT and TM results against each other. In addition, these contexts can be compared to already extracted RDFS triplet contexts that are stored in a separate language resource, which currently is further investigated in an PhD research (Weissgerber, forthcoming).

Thirdly, we are optimistic that the further evolution of MT will foster and increase the success of our MT/TM employment environment, that the adopted engineering approach will further mature, and that additional community exchange and collaboration will eventually contribute to its overall success.

In the next Section, we will discuss the enormous benefits of sharing language resources and the impact on MT.

4.   MT's Future in Glocalization

We have to distinguish various types of language resources, which we call different species: these can be raw unstructured or semi-structured data, e.g. XML documents, structured data from a database system, enriched data such as linguistically tagged data with grammatical and semantic information, aligned multilingual data of a translation memory system or a terminology management system, and many many others, as well as software applications and tools that technically facilitate data and information exchange, optimization, correlation, interrelationship, and so forth. An MT system or MT service is one very specific incarnation of such a software application. Therefore, when we talk about language resources in the following sections, MT is one particular species in a Language Resources Ecosystem (LRE).

---

[4] A locally installed rule-based MT system would certainly benefit from an application-specific user lexicon in these cases. However, we wanted to elaborate the potentials of future cloud-based automated translation services.

[5] RDF(S) stands for Resource Description Format (Schema), and belongs to the series of W3C recommendations for the Semantic Web.

4.1 Sharable Language Resources

The sharability of these language resources is an essential and critical factor to administer, master and control the new, highly proactive, global communications demands and needs of individuals and businesses in a multitude of applications that need specific language information to accomplish their tasks in a sustainable and responsible style.

Sharing in this context means that the resources are employable and deployable in and across different applications, and that they are widely and translucently distributed with documented interfaces (APIs, SDKs, etc.). Applications may range from standard technical communication and documentation of global enterprises through user generated Web content, such as blogs, podcasts, etc., to interpersonal communication, for example, on a foreign fish market or a medical consultation with a smartphone.

In general, enablers of sharability are standardized or agreed upon descriptions, formats and protocols such as database schemata, exchange and interface specifications, and so forth.

The active sharing of language resources has certainly an economic dimension, and today, advances in technology and the Internet allow many language resources to be produced for more or less nothing because they can be successfully extracted and discovered in the huge amounts of information available in multiple languages with sophisticated engineering approaches, statistical means and collaborative efforts. The questions here is "Who owns the data that we can discover and extract on the Internet?" See for example the recent discussions and debates on Google's data hunger, their data mining activities, and their various approaches to gain access to IP protected data as well as their data storage and applications regimes and policies.

To offer language resources for free is therefore an eventual consequence of this situation and is determined by two important factors that are more valuable than money: the popular reputation of what is on offer, and the time we have available for it to invest in, for example, crowdsourcing efforts. This then goes in-line with the change of the old money-centric economy to an economy of reputation and of time which we are facing right now.

Nevertheless, besides offering language resources as a free good – free language resources –

there is also the possibility to make them available for everyone with certain restrictions such as an open source license or a Creative Commons license – open language resources – and there is the possibility to make them only available to a specific group of sharers as this is already demonstrated by, for example, the TAUS Data Association – members only language resources.

In each case, we have to care about aspects of quality, security, trust, privacy and fitness for use of the offered language resources, which need supplementary investments. These aspects open yet another perspective of sharability, which is beyond the simple distribution at no costs at all because of the digital availability of the resources, language resources producers have responsibility and may offer additional services for and with their resources. The quality of these services then is more than the resources' actual deployability and overall usability, and cannot be determined by download rates and the reputation the resources gain in different communities and with their applications. However, such additional services, on the one hand, may bear the danger of a possible lock-in, but on the other hand, they might also be a matter of differentiation which is the standard evolutionary emergence of innovations.

4.2 Global Language Services

Written and spoken language are human's most important information exchange carriers, therefore language resources as discussed in the previous section will be needed in any "humanized" application across domains, cultures and experiences. Since we are no longer either information producers or information consumers but information prosumers, new ways of human-computer interaction will emerge. The added-value of the accompanying application services will be quality as a real user experience and the recognition of a continuous flow in time as well as portability and transparency combined with general responsibility and sharability.

Possible products and applications will appear in every day situations such as personal communications and in specific situations such as medical consultation and general health care, nutrition advice, efficient energy consumption advice, and so forth, and for highly adaptable machine translation to account for cross-language and transcultural needs.

These applications are in average accessible from smart mobile devices that are connected to the Internet to gain on-demand computational power and ambient language resources based on purpose and location mainly through cloud computing based services. Payment modes for these services will be either on a subscription base or by micro payment and always in a pay-per-use manner which we know from our water, gas or electricity suppliers.

Automation and ambient adaptability are key to the success of global language services that have to offer process modeling capabilities, control and monitoring facilities as well as open interfaces to allow for usable connectability and sharability.

5. Conclusions and Perspectives

Although we have painted the emerging era of glocalization with different sharable language resources in various facets with a broad brush, it should be obvious that the future is essentially in integrating these resources into various transparent services that add value and quality as an entire user experience. In particular, these services must also facilitate privacy and security of the individually optimized resources as well as their accessibility and sharability.

Because of the enhanced network effect of all things being digital and some apparently free at all, it is not easy for any language resource provider to become profitable with advertising – Google and some few others have already conquered this domain – but thorough fee-based services with built-in ambient adaptability will compensate the cost for provision, maintenance and enhancements so that language ecosystems sustain and further evolve.

In the next years, self-adapting and self-learning machine translation services of different kinds including sophisticated post-editing support services will play an important role in solving the increasing multilingual transcultural communications needs and demands. For these services sharable language resources across domains and applications are extremely necessary for their survival.

Because the digital network architecture as it exists today naturally incubates monopolies, responsible prosumers will resist these trends and actively support the building of a real sustainable network (Sorenson, 2010).

6. References

Andrä, Sven C., and Jörg Schütz. 2009a. *MT Bazaar: Translation Ecosystems in the Cloud. Machine Translation Summit XII,* Ottawa, Canada.

Andrä, Sven C., and Jörg Schütz. 2009b. *Cloud Computing in GILT Ecosystems and Evolution. ASLIB Conference Translating and the Computer 31*, London, UK.

Andrä, Sven C., and Jörg Schütz. 2010. *Internationalization meets Semantic Technologies. Worldware Conference 2010*, San José, California, USA.

Asia Online. 2009. *Language Studio Enterprise – API Reference Documentation*. Singapore and Bangkok, Thailand.

Banerjee, Satanjeev. and Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* In: Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL 2005), Ann Arbor, Michigan, USA.

Callison-Burch, Chris, Miles Osborne and Philipp Koehn. 2006. *Re-evaluating the Role of BLEU in Machine Translation Research*. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).

Koen, Philipp. 2010. *Moses – Statistical Machine Translation System, User Manual and Code Guide.* Edinburgh, UK.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: A method for automatic evaluation of machine translation. In: Proceedings of ACL 2002.*

Schütz, Jörg. 2008. *Artificial Cognitive Post-Editing Intelligence. Eighth AMTA Conference,* Waikiki, Hawai'i, USA.

Schütz, Jörg. 2001. *Blueprint for MT Evolution: Reflections on Elements of Style. Workshop "MT Roadmap 2010" Machine Translation Summit VIII,* Santiago de Compostella, Spain.

Sorensen, Sarah. 2010. *The Sustainable Network.* O'Reilly Media, Sebastopol, California, USA.

TAUS. 2010. *TAUS predicts.* Webinar http://www.translationautomation.com/events/webinars/taus-predicts.html.

Weissgerber, Alexandra. Forthcoming. *Semantically Enriched Business Process Modeling*. PhD Dissertation, Saarland University, Saarbrücken, Germany.