Evaluation of SMT in localization to under-resourced inflected language

Raivis Skadiņš, Maris Puriņš, Inguna Skadiņa, Andrejs Vasiļjevs
Tilde SIA
Vienibas gatve 75A, Riga
Latvia, LV-1004
{raivis.skadins,maris.purins,inguna.skadina,andrejs}@tilde.lv

Abstract

Although Machine Translation is very popular for personal tasks, its use in localization and other business applications is still very limited. The paper presents an experiment on the evaluation of an English-Latvian SMT system integrated into SDL Trados which has been used in an actual localization assignment by a professional localization company. We show that such an integrated localization environment can increase the productivity of localization by 32.9% without a critical reduction in quality.

1 Introduction

The rapid advancement of machine translation technologies, especially statistical machine translation (SMT), draws attention to questions of its usability in practical, business-oriented applications.

One promising area where industry could benefit from SMT is localization. Translation memories (TM) have been in use in localization for more than 10 years to increase productivity. Translation memories can significantly improve the efficiency of localization if the new text is similar to the previously translated material. However, if the text is in a different domain than the TM or in the same domain from a different customer using different terminology, support from the TM is minimal. For this reason the localization industry is increasingly interested in combining translation memories with machine translation solutions adapted for the particular domain or customer requirements.

For the development of MT in the localization and translation industry, huge pools of parallel texts in a variety of industry formats have been accumulated, but the use of this data alone does not fully utilize the benefits of modern MT technology. At the same time, this industry experiences a growing pressure on efficiency and performance, especially due to the fact that volumes of texts that need to be translated are growing at a greater rate than the availability of human translation, and translation results are expected in real-time.

Increasing the efficiency of the translation process without a degradation of quality is the most important goal for a localization service provider.

In this paper we present an experiment on the application of an English-Latvian SMT in localization through the integration of MT into the SDL Trados 2009 translation environment. Similarly to Plitt and Masselot (2010) we measure performance of a translator translating with and without MT. In addition a quality assessment for texts was performed according to the standard internal quality assessment procedure.

2 Related work

Although the idea to use MT in the localization process is not new, it has not been explored widely. Different aspects of post-editing and machine translatability have been researched since the 90-ies. A comprehensive overview of research on machine translatability and post-editing has been provided by O´Brien (2005). She also analyses the potential of a keyboard-monitoring program and Choice Network Analysis for measuring the effort involved in post-editing MT output. However this work mainly

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 35–40
Leuven, Belgium, May 2011

concentrates on the cognitive aspects, not so much on productivity in the localization industry.

Recently several productivity tests have been performed in translation and localization industry settings at Microsoft (Schmidtke, 2008), Adobe (Flournoy and Duran, 2009) and Autodesk (Plitt and Masselot, 2010).

The Microsoft Research trained SMT on MS tech domain was used for 3 languages for Office Online 2007 localization: Spanish, French and German. By applying MT to all new words on average a 5-10% productivity improvement was gained.

Adobe performed two experiments. At first small test set of 800-2000 words was machine translated and post-edited. Then, based on the positive results, about 200,000 words of new text were localized. The rule-based MT was used for translation into Russian (PROMT) and SMT for Spanish and French (Language Weaver). Initial results were very positive implying that "a translator's daily output can be produced by a post-editor in less than two hours". However, after detailed investigations authors report the speed-up between 22% and 51%.

At Autodesk, a Moses SMT system was evaluated for translation from English to French, Italian, German and Spanish by three translators for each language pair. To measure translation time a special workbench was designed to capture keyboard and pause times for each sentence. Authors reported that although by using MT all translators worked faster, it was in varying proportions: from 20% to 131%. They concluded that MT allowed translators to improve their throughput on average by 74%. They also reported that optimum throughput has been reached for sentences of around 25 words in length.

3 Localization scenario with the LetsMT! platform

The current experiment was performed in the scope of the LetsMT! project[1] and using tools and a platform developed in the project. The aim of LetsMT! (Vasiljevs et al., 2010) is to exploit the huge potential of existing open SMT technologies by developing an online collaborative platform for data sharing and MT building. This platform will support uploading of public and proprietary MT training data and building of multiple MT systems by combining and prioritizing data.

LetsMT! services are focused on two application scenarios –use in localization and translation and for online translation of financial news.

The goal of the online MT service for the localization industry is to increase the efficiency of work performed by industry professionals – localization and translation service providers (LSPs), organizations with multilingual translation needs, and freelance translators. LetsMT! services will support generation of customized MT from user data. The project has the specific aim to facilitate MT development and application for highly inflected languages such as Latvian, Lithuanian, Estonian, Czech, Slovak, Polish, Croatian.

The initial collection of corpora is focused on parallel texts in these languages and English in IT and Telecommunication domain. Once the service is released, the range of sub-domains and languages supported will be largely user-driven, i.e., determined by the requirements and opportunities in the localization and translation market.

This service will enable professional users to generate and employ customized MT services of higher quality based on specific terminology and style required by their clients. It will take into account the workflow, technical requirements and legal ramifications characteristic of the localization industry.

The Giza++ and Moses SMT toolkits (Koehn et al., 2007) are used for data alignment, training of SMT models and translation (decoding). SMT platform is integrated with professional translation tools to enable smooth work of translators in familiar environment.

4 Evaluated SMT system

For training the SMT systems, both monolingual and bilingual sentence-aligned parallel corpora of substantial size are required.

The parallel training corpus includes publicly available DGT-TM[2] (1.06 M sentences) and OPUS EMEA (0.97 M sentences) corpora (Tiedemann, 2009), as well as a proprietary localization corpus (1.27 M sentences) obtained from translation memories that were created during the localization of interface and user assistance materials for software and user manuals for IT&T appliances. To increase word coverage we included word and phrase translations from bilingual dictionaries (0.51 M units). This parallel data come from reliable sources and is of high

[1] http://www.letsmt.eu

[2] http://langtech.jrc.it/DGT-TM.html

quality. We also used a larger selection of but not as reliable parallel data automatically extracted from comparable web corpus. These parallel data are extracted from c.a. 159,000 comparable html and pdf documents crawled from the web (0.9 M sentences) and from 104 works of fiction (0.66 M sentences). The total size of the English-Latvian parallel data used to train the translation model was 4.1 M sentence pairs.

The monolingual corpus was prepared from news articles from the Web and the monolingual part of the parallel corpora. The total size of the Latvian monolingual corpus was 391 M words.

We used the Moses SMT toolkit for SMT system training and decoding. The SMT system was extended within the Moses framework by integrating morphologic knowledge (Skadiņš et al., 2010). Latvian belongs to the class of highly inflected languages with a complex morphology. There are over 2000 different morphology tags for Latvian. The high inflectional variation of target language increases data sparseness at the boundaries of translated phrases, where a language model over surface forms might be inadequate to estimate the probability of target sentence reliably. Following the approach of English-Czech factored SMT (Bojar et al., 2009) we introduced an additional language model over disambiguated morphologic tags in the English-Latvian system. The tags contain morphologic properties generated by a statistical morphology tagger. The order of the tag LM was increased to 7, as the tag data has significantly smaller vocabulary.

The evaluation and development corpora were prepared separately. For both corpora we used the same mixture of different domains and topics representing the expected translation needs of a typical user in the general domain. The development corpus contains 1000 sentences, while the evaluation set is 500 sentences long. Both corpora contain information technology (IT) domain texts (c.a. 20%).

We used the BLEU (Papineni et al., 2002) metric for automatic evaluation. The BLEU score of the SMT system is 35.0.

We used the SDL Trados 2009 CAT environment to evaluate the use of the MT system in the localization scenario. The MT system is running on LetsMT! platform and is accessible using a web service interface based on the SOAP protocol.

To integrate the SMT system in SDL Trados we developed a plug-in using standard MT integration approach described in SDL Trados SDK.

5 Evaluation task

Evaluation in the localization scenario was based on the measurement of translation performance. Performance was calculated as the number of words translated per hour.

The evaluation was made in the software localization domain.

5.1 Evaluation scenarios

For the evaluation two test scenarios were employed: (1) a baseline scenario with TM only and (2) an MT scenario with a combination of TM with MT. The baseline scenario established the productivity baseline of the current translation process using SDL Trados Studio 2009 when texts are translated unit-by-unit (sentence-by-sentence). The MT scenario measured the impact of using MT in the translation process when translators are provided not only matches from the translation memory (as in baseline scenario), but also MT suggestions for every translation unit that does not have 100% match in translation memory. Suggestions coming from the MT were clearly marked (see Figure 1).

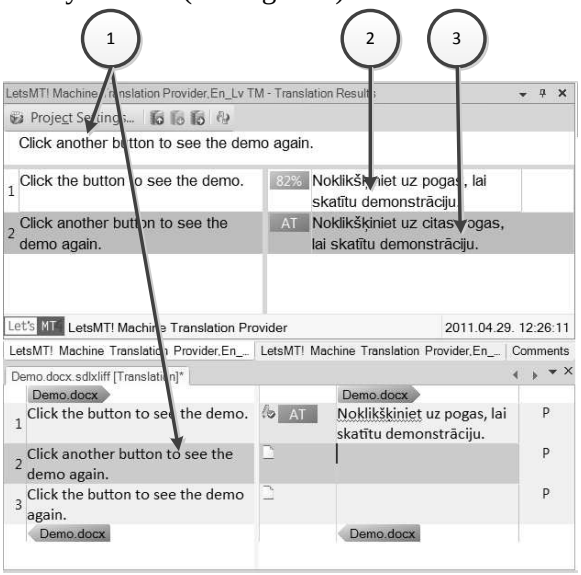

Figure 1. Translation suggestions in SDL Trados Studio 2009; 1 – a source text, 2 – a suggestion from the TM, 3 – a suggestion from the MT.

We chose to mark MT suggestions clearly because it allows translators to pay more attention to these suggestions. Typically translators trust to suggestions coming from the TM and they make only small changes if it is not a 100% match. Translators are not double-checking terminology, spelling and the grammar of TM suggestions, because the TM contains good quality data. But translators must pay more attention to sugges-

tions coming from MT, because MT output may be inaccurate, ungrammatical, it may use the wrong terminology etc.

In both scenarios translators were allowed to use whatever external resources needed (dictionaries, online reference tools etc.), just as during regular operations.

Five (5) translators with different levels of experience and average performance were involved in the evaluation.

The first translation of each translator in the MT scenario was removed from the results analysis to avoid "start-up" impact.

5.2 Evaluation of translation quality

The quality of each translation was evaluated by a professional editor based on the standard quality assurance process of the service provider. The editor was not made aware whether the text was translated using the baseline scenario or the MT scenario. An error score was calculated for every translation task. The error score is a metric calculated by counting errors identified by the editor and applying a weighted multiplier based on the severity of the error type. The error score is calculated per 1000 words and it is calculated as:

$$ErrorScore = \frac{1000}{n} \sum_i w_i e_i$$

where

- n is a number of words in a translated text,
- e_i is a number of errors of type i,
- w_i is a coefficient (weight) indicating severity of type i errors.

There are 15 different error types grouped in 4 error classes – accuracy, language quality, style and terminology. Different error types influence the error score differently because errors have a different weight depending on the severity of error type. For example, errors of type *comprehensibility* (an error that obstructs the user from understanding the information; very clumsy expressions) have weight 3, while errors of type *omissions/unnecessary additions* have weight 2.

Depending on the error score the translation is assigned a translation quality grade: Superior, Good, Mediocre, Poor and Very poor (Table 1).

5.3 Test set

The test set for the evaluation was created by selecting documents in the IT domain from the tasks that have not been translated by the translators in the organization before the SMT engine was built. This ensures that translation memories do not contain all the segments of texts used for testing.

Table 1. Quality evaluation based on the score of weighted errors

Error Score	Quality Grade
0…9	Superior
10…29	Good
30…49	Mediocre
50…69	Poor
>70	Very poor

Documents for translation were selected from the incoming work pipeline if they contained 950-1050 adjusted words each. Each document was split in half and the first part of it was translated as described in the baseline scenario and the second half of the document – using the MT scenario. The project manager ensured that each part of a single document was translated by different translators so the results are not affected by due to translating a familiar document.

Altogether 54 documents were translated for the English -> Latvian language pair. Each document tagged for evaluation was entered in the translation project tracking system as a separate translation task. An adjusted word is a metric used for quantifying work to be done by translators. Larger documents were split into the several fragments.

Although a general purpose SMT system was used, it was trained using specific vendor translation memories as a significant source of parallel corpora as described in the previous sections. Therefore the SMT system may be considered slightly biased to a specific IT vendor, or a vendor specific narrow IT domain. The test set contained texts from this vendor and another vendor whose translation memories were not included in the training of the SMT system. We will call these texts as *in narrow IT domain* and *in broad IT domain* for easier reference in the following sections. Approximately 33% of texts translated in each scenario where *in broad IT domain*.

6 Results

The results were analyzed for 46 translation tasks (23 tasks in each scenario) by analyzing average values for translation performance (translated words per hour) and error score for translated texts.

Usage of MT suggestions in addition to the use of the translation memories increased productivity of the translators in average from

550 to 731 words per hour (32.9% improvement). There were significant performance differences in the various translation tasks; the standard deviation of productivity in the baseline and MT scenarios were 213.8 and 315.5 respectively.

At the same time the error score increased for all translators. Although the total increase in the error score was from 20.2 to 28.6 points, it still remained at the quality evaluation grade "Good".

Grouping of the translation results by narrow/broad domain attribute reveals that MT-assisted translation provides better increase in translation performance for narrow domain (37%) than for broad domain texts (24%). Error scores for both text types are very similar 29.1 and 27.6, respectively.

Grouping of errors identified by error classes reveal the increase of number of errors shown in Table 2.

Table 2. Comparison by error classes (error score)

Error Class	Baseline scenario	MT scenario
Accuracy	6	9
Language quality	6	10
Style	3	4
Terminology	5	7

There were significant differences in the results of different translators from performance increase by 64% to decreased performance by 5% for one of the translators.

Analysis of these differences requires further studies but most likely they are caused by working patterns and the skills of individual translators. Study and dissemination of the best practice of the most efficient translators would be needed to get the most benefits from MT application.

7 Conclusions

This is one of the first evaluations of MT for less-resourced highly inflected language in the localization environment. The results of our experiment clearly demonstrate that it is feasible to integrate the current state of the art SMT systems for highly inflected languages into the localization process.

The use of the English->Latvian SMT suggestions in addition to the translation memories in the SDL Trados CAT tool leads to the increase of translation performance by 32.9% while maintaining an acceptable quality of the translation. Even better performance results are achieved when using a customized SMT system that is trained on a specific domain and/or same customer parallel data.

Error rate analysis shows that overall usage of MT suggestions decrease the quality of the translation in all error categories, but especially in language quality. At the same time this degradation is not critical and the result is acceptable for production purposes.

Acknowledgements

The research within the LetsMT! project leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web, grant agreement no250456.

References

Bojar O., Mareček D., Novák V., Popel M., Jan Ptáček J., Rouš J., Žabokrtský Z. 2009. English-Czech MT in 2008. *Proceedings of the Fourth Workshop on Statistical Machine Translation. Association for Computational Linguistics*, Athens, Greece.

Flournoy, Raymond and Christine Duran. 2009. Machine translation and document localization at Adobe: From pilot to production. *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, Ottawa, Canada.

Koehn P., Federico M., Cowan B., Zens R., Duer C., Bojar O., Constantin A., Herbst E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, *in Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, 177-180.

O´Brien, Sharon. 2005. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, 19(1):37–58, March 2005.

Papineni K., Roukos S., Ward T., Zhu W. 2002. BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL).*

Plitt Mirko, François Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93(January 2010): 7–16

Schmidtke, Dan. 2008. Microsoft office localization: use of language and translation technology. *URL http://www.tm-europe.org/files/resources/TM-Europe2008-Dag-Schmidtke-Microsoft.pdf.*

Skadiņš, Raivis, Kārlis Goba and Valters Šics. 2010. Improving SMT for Baltic Languages with Fac-

tored Models. *Proceedings of the Fourth International Conference Baltic HLT 2010*, Riga.

Tiedemann, Jörg. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds.) Recent Advances in Natural Language Processing* (vol V), John Benjamins, Amsterdam/Philadelphia, 237-248.

Vasiljevs, Andrejs, Tatiana Gornostay and Raivis Skadins. 2010. LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. *Proceedings of the Fourth International Conference Baltic HLT 2010*, Riga.