Real-world challenges in application of MT for localization: the Baltic case

Mārcis Pinnis marcis.pinnis@tilde.lv
Raivis Skadiņš raivis.skadins@tilde.lv
Andrejs Vasiļjevs andrejs@tilde.com
Tilde, Riga, LV-1004, Latvia

Abstract

In this paper we share our experience from implementing machine translation in localization into relatively small languages of the three Baltic countries – Latvian, Lithuanian, and Estonian. We describe our approach in improving terminology translation and consistency by preprocessing of the source text and performing term integration. We present results of a formal evaluation of MT impact on the productivity of translators' work. Quality, usability, increase of productivity of translation process has been evaluated. Our evaluation shows that the use of the SMT suggestions in addition to the translation memories in the SDL Trados CAT tool leads to the increase of translation performance by up to 32.9% while maintaining an acceptable quality of translation.

1. Introduction

The localization industry experiences increased pressure to provide cheaper and faster services, particularly for the languages of smaller countries where return of investments is much lower than from the larger markets. This motivates localization service providers to look for tools that can boost productivity and streamline the translation process.

In the last few years machine translation has been increasingly integrated into the day-to-day localization process at numerous providers of translation and localization services for larger languages. In this paper we share our experience from implementing machine translation in localization into relatively small languages of the three Baltic countries – Latvian, Lithuanian, and Estonian. This paper provides an outlook of the Tilde work in this area published in several research papers (Pinnis et al., 2013b; Pinnis & Skadiņš, 2012; TaaS, 2014; Skadiņš et al., 2014).

We describe our approach in improving terminology translation and consistency by preprocessing of the source text and performing terminology integration in SMT systems. Specific attention is devoted to treating of multiword units and inflectional forms of terms.

We are presenting our work on evaluation of applicability of state-of-the-art MT systems for these languages in the real-world localization. We analyze, compare and discuss assessment results of the typical problems in translation into Baltic languages. We are discussing how inflectional nature of languages and limited number of parallel data lead to such frequent MT errors as incorrect word forms, untranslated words and missing words, and we propose strategies how to mitigate these problems by adapting translation engines to particular domain or project. We describe integration of MT into translation workflow in pre-translation scenario and in a run-time usage with segment-by-segment translation through a plug-in for popular

CAT tools like SDL Trados and Kilgray memoQ. We also analyze issues faced in a translation of segments with in-line tags and present our approach in their processing.

We present results of a formal evaluation of MT impact on the productivity of translators' work. Quality, usability, increase of productivity of translation process has been evaluated. The evaluation process involved measurement of productivity, inspection of translations, and classifying errors according to the Quality Assessment (QA) matrix with 15 different error types grouped in 4 error classes: accuracy, language quality, style, and terminology.

2. Handling of Terminology Translation

This section summarizes our work towards development of term-aware SMT systems previously published in Pinnis et al. (2013b), Pinnis & Skadiņš (2012), and TaaS (2014).

2.1. Problem

Correct and consistent handling of terminology is a very important requirement in localization. It is also one of key features that is assessed when performing manual quality evaluation of translations in professional translation and localization services according to the QA Model of the Localization Industry Standards Association (LISA). The QA form for translations in the Baltic localization service provider Tilde is based on the LISA (2008) QA Model. Therefore, it is essential that any automated translation solution (if it is to be introduced into professional translation service workflows) provides support for correct and consistent handling of terminology.

Currently the dominant machine translation paradigm is the phrase-based statistical machine translation. However, current SMT phrase-based models, including Moses (Koehn et al., 2007), do not handle terminology translation. Although domain adaptation can be performed using additional in-domain training data (Koehn and Schroeder, 2007), such an approach is very resource intensive as it requires gathering of the resources (parallel and monolingual corpora) for each individual domain and for smaller projects or for languages with limited resources this is not an option. This makes terminology integration with the standard approaches expensive (in terms of time) and for less resourced languages in many cases also not feasible (due to lack of parallel or monolingual in-domain corpora). Several of the main issues of terminology translation without explicit support for terminology integration within SMT systems are as follows:

- Terms may be translated using wrong translation equivalents. That is, the translation equivalents may be: 1) from a different domain, 2) from obsolete variants of terms, 3) in abbreviated or non-abbreviated forms (contrary to the terms' form in the source language), 4) used by a client's competitor, etc. For example, the term „*tablet*" is ambiguous – it can refer to a popular consumer electronics product (a tablet computer), a number of sheets of paper fastened together along one edge (according to WordNet 3.1[1]), a pill used in medicine, and others. An SMT system would translate this term in every single case according to its statistical translation and language models. In other words, a term would be translated using the most probable phrase alignment, which may not be the correct one.
- Terms may be split into several parts during translation. This problem may occur because SMT systems (including the Moses system) use reordering (also known as distortion) models that allow reordering translated phrases in the target language.

[1] More information on WordNet can be found online at: http://wordnet.princeton.edu/.

- Terms may also be missing in the SMT system's models, which means that out-of-vocabulary terms would not be translated.
- Multi-word terms may also be translated by breaking morpho-syntactic agreements between constituents of terms. For instance, when translating into morphologically rich languages (e.g., Latvian, Estonian, Czech, etc.) it is important that morpho-syntactic agreements are modelled (or at least correctly transferred) in the target language. When translating into Latvian it is important, for instance, for adjectives to be generated with the same gender, number, and case properties as the head noun in the immediate noun phrase the adjectives belong to.
- When localizing software or translating documents from specific clients, the clients may request using their specific terminology. SMT systems rely on statistics when translating terms and not on pre-defined term collections. For instance, if we have a term collection from a client that specifies that the term "*web service*" has to be translated as "*tīmekļa pakalpe*" in Latvian, then the SMT system should be able to support such user requests. However, current SMT solutions do not offer such functionality.

Researchers have previously tried to address the terminology translation quality issues with different methods that allow integrating bilingual term collections in SMT systems during SMT system training. For instance, Bouamor et al. (2012) have observed a gain of +0.3 BLEU (Papineni et al., 2002) points for French-English SMT by simply adding automatically extracted bilingual terminology to the parallel corpora used for SMT system training.

Several research works are focused towards providing runtime integration of term collections provided by users into SMT systems. For instance, the popular Moses SMT platform supports translation of pre-processed documents where explicit translations of terms can be marked using XML tags. Carl and Langlais (2002) in their research showed that using term collections in such a way could increase the translation performance for the English-French language pair. Babych and Hartley (2003) showed that for NEs (namely, organization names) special "*do-not-translate*" lists improved translation quality for the English-Russian language pair using a similar pre-processing technique that restricts translation of identified phrases. However, such approaches have been investigated either for languages with simple morphology or categories of phrases that are rarely translated or even left untranslated (e.g., many company and organization names). A recent study in the FP7 project TTC (2013) has shown that for English-Latvian the pre-processing does not yield positive results for term translation. Hálek et al. (2011) also showed that the translation performance with on-line pre-processing drops according to BLEU for English-Czech named entity translation. This proves that the method is not stable when translating into morphologically rich languages, that is, languages with a high level of inflection (e.g., the Baltic and Slavic languages).

Terminology integration in SMT systems can be also achieved through domain adaptation with the help of in-domain parallel and monolingual corpora. However, acquisition of in-domain parallel corpora may be unfeasible in many situations. SMT system adaptation with the help of language models trained on in-domain monolingual data in addition to an out-of-domain language model (which may be trained on much larger data sets), on the other hand, has shown to be an effective method (Koehn and Schroeder, 2007; Lewis et al., 2010) for tailoring SMT systems to a specific domain. Although SMT domain adaptation has been an active field in the machine translation research community, the majority of practical SMT applications rely solely on collecting large amounts of domain specific corpora. Moreover, there are not so many even more advanced solutions, which would focus on special handling of terminology. It is assumed that training data will contain translations with terminology and SMT will learn accurate terminology from training data. However, it is not usually the case as training data, even if it is in the same domain, can contain contradicting terminology – industry or corporate specific synonyms in product- or vendor-biased terminology.

2.2. Approach

In this paper we summarise our methods for effective bilingual terminology integration in SMT systems during SMT system training. We assume that for SMT system training we have all the necessary data already available (a parallel corpus, a monolingual corpus, and an in-domain bilingual term collection). The term collection may be created manually by a translator or terminologist, it can be acquired fully automatically from parallel or comparable corpora (for instance, as shown by Pinnis et al., 2012), or it can be created in a semi-automated fashion from monolingual documents in the required domain with the help of the TaaS platform (Pinnis et al., 2013a).

Similarly to related research (Bouamor et al., 2012), our methods for terminology integration in SMT systems during system training require that the bilingual term collections are added to the parallel corpus, which is used for translation model training, and the target language terms are added to the monolingual corpus, which is used for language model training. This method, although being very simple, is quite efficient, because it ensures that the terms that are not covered by both the parallel corpus and the monolingual corpus (i.e., terms that can be considered as out-of-vocabulary terms) will have at least one translation hypothesis. Obviously, for languages that feature rich morphologies this method won't be productive if terms in term collections will be stored in their canonical forms. Nevertheless, this method can be efficient in the following three scenarios: 1) when translating from and to languages with little morphological inflection (e.g., from or to English, German, French, etc.), 2) when using term collections acquired in an automatic process from, e.g., parallel data or comparable data (where terms are already stored in surface forms that are common in different contexts), and 3) when training SMT systems, because even if term pairs are provided in canonical forms, they can indirectly help improving word alignment and subsequently also phrase alignment quality.

Further, we transform the Moses phrase table of the translation model into an in-domain term-aware phrase table. This is performed by adding a 6th feature to the default five features that are used in Moses phrase tables. The 6th feature receives the following values:

- "1" if a phrase on both sides (in both languages) does not contain a term pair from a bilingual term collection. If a phrase contains a term only on one side (in one language), but not on the other, it receives the value "1" as such situations may indicate that the given phrase pair contains an out-of-domain translation equivalent.
- "2.718" if a phrase on both sides (in both languages) contains a term pair from the bilingual term collection.

In order to identify whether a phrase in the Moses phrase table contains a given term or not (also in different inflected forms), phrases and terms are stemmed prior to comparison. After tuning of the SMT system with Minimum Error Rate Training (MERT; Bertoldi et al., 2009), the 6th feature allows assigning higher translation probabilities to in-domain translation hypotheses, thereby decreasing the possibility to select an out-of-domain translation of terms and at the same time increasing the overall translation quality.

2.3. Evaluation

To show the potential of the terminology integration methods, an evaluation in a translation task of car service manuals was performed. All SMT systems were trained using the LetsMT SMT platform (Vasiļjevs et al., 2012). In total, two scenarios were analysed. In the first scenario the baseline (out-of-domain) SMT system was trained using the DGT-TM (Steinberger et al., 2012) English-Latvian parallel corpus (release of 2007). The corpus consists of approximately 804 thousand unique parallel sentence pairs. In the second scenario the baseline system was trained using a much larger corpus of 5.36 million unique parallel sentence pairs consisting of publicly available and proprietary corpora. All SMT systems were tuned and evaluated using 1,745 and 872 unique sentence pairs respectively from a small proprietary in-domain parallel corpus. For domain adaptation, in the first scenario an in-domain monolingual corpus of 271 thousand unique Latvian sentences collected from the Web (Pinnis et al., 2012) was used to train an in-domain language model. In both scenarios one bilingual term collection of 979 term pairs was used for terminology integration in the SMT systems. Automatic evaluation results using different automatic evaluation methods (NIST (Doddington, 2002), BLEU, METEOR (Denkowski & Lavie, 2011), and TER (Snover et al., 2006)) are given in Table 1 (first scenario with the small corpora systems) and Table 2 (second scenario with the large corpora systems).

System	BLEU	BLEU (CS)	NIST	NIST (CS)	TER	TER (CS)	METEOR	METEOR (CS)
Baseline	10.97	10.31	3.9355	3.7953	89.75	90.40	0.1724	0.1301
+ in-domain LM	11.30	10.61	3.9606	3.8190	89.74	90.34	0.1736	0.1312
+ terms	13.50	12.65	4.2927	4.1105	88.86	89.70	0.1878	0.1443
+ 6th feature	**13.61**	**12.78**	**4.3514**	**4.1747**	**88.54**	**89.32**	**0.1920**	**0.1469**

Table 1. English-Latvian automotive domain SMT system adaptation results using DGT-TM

System	BLEU	BLEU (CS)	NIST	NIST (CS)	TER	TER (CS)	METEOR	METEOR (CS)
Baseline	15.85	15.00	4.8448	4.6934	73.80	75.12	0.2098	0.1651
+ terms	17.24	16.12	5.0020	4.8278	72.16	73.59	0.2163	0.1717
+ 6th feature	**18.21**	**17.08**	**5.1476**	**4.9626**	**70.22**	**71.62**	**0.2191**	**0.1747**

Table 2. English-Latvian automotive domain large SMT system adaptation results

The results show that by simply adding terms to the training data allows achieving significant translation quality improvements over baseline systems in both scenarios. As in the second scenario, the term out-of-vocabulary rate before the addition is lower, the SMT quality improvement is also lower. The results also show that the transformation of the SMT system phrase tables into term-aware phrase tables allows boosting the SMT quality even further reaching a cumulative SMT quality improvement of 24% for the smaller system and approximately 15% for the larger system.

3. Addressing morphological complexity

The languages of the Baltic States belong to the class of inflected languages with complex morphology and rather free word order, which makes them complicated subjects for statistical MT (Koehn et al., 2009). The lack of necessary language technologies and the need for large amounts of parallel corpora make MT even more difficult. According to a recent report from META-NET, the languages of the Baltic States are at risk of digital extinction and MT technologies are weakly developed for them (Rehm & Uszkoreit, 2012). At the same time there have been numerous academic and industrial activities to research and build MT systems.

To train our SMT systems for application in localization we use the LetsMT platform (Vasiļjevs et al., 2012). When training general domain SMT systems on a large parallel corpora, we see that even without any language specific adaptation a standard phrase-based approach can result in a relatively good quality MT. To further increase quality of MT systems we can integrate language pair specific methods. The most promising method to incorporate linguistic knowledge in SMT is to use morphology in factored SMT models. Our experiments show that an additional language model over morpho-syntactic tags improves inter-phrase consistency (Skadiņš et al. 2010) when translating to morphology rich language. In translation from Baltic language, quality can improved by pre-processing of the source text to reduce the morphological complexity and to address the data sparseness (Deksne & Skadiņš, 2012).

At the same time we have to note that these methods show relatively modest quality boost in comparison to the improvement achieved by multiplying the amount of parallel training data. Still they are quite useful as for smaller languages parallel data is a very limited resource.

To get better results from existing data, our current work is focused on improving of word alignment by implementing calculation over lemmas instead of the surface forms.

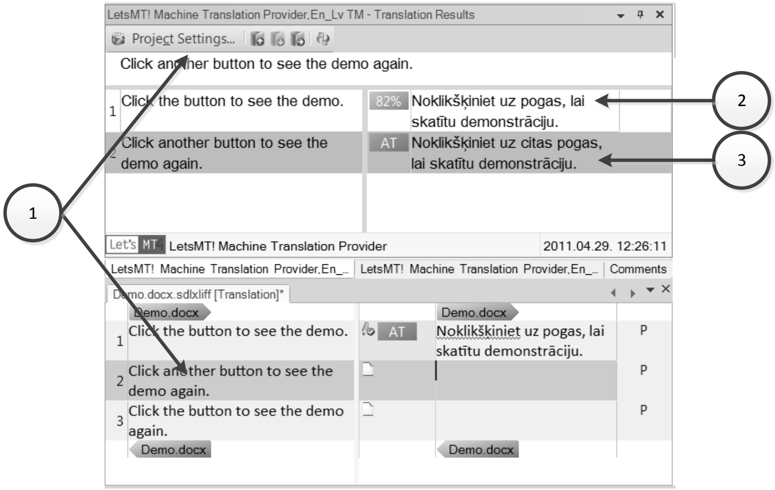
4. Integration of MT in productivity tools

Easy integration of the MT solution in CAT tools is very important. The LetsMT platform provides several integration scenarios:

- Segment-by-segment translation using a CAT tool plug-in,
- Pre-translation of a document or project using a CAT tool plug-in,
- File pre-translation using online MT functionality on the Tilde MT website.

The first scenario allows using MT in the same manner as a translator uses translation memories (TM). A translator is translating a document segment by segment and receives translation suggestions both from translation memories and the MT system. For each segment, translator can decide whether to use and correct the provided suggestion or to translate the segment from scratch. In the CAT tool translation suggestions from a selected MT system are provided to the translator as shown in the Figure 1. This is not a pure MT post-editing scenario, because the translator is not asked to post-edit the MT output. Instead MT output is provided as a possibly useful suggestion. We chose to clearly mark MT suggestions as they need additional attention of translators. Usually translators trust suggestions coming from the TM and they apply only minor changes if necessary. They often do not double-check terminology, spelling, and grammar, because the TM is supposed to contain good quality data. However, translators must pay more attention to suggestions coming from MT, because MT output may be inaccurate, ungrammatical, it may use wrong terminology, etc.

Currently the Tilde MT provides CAT tool plug-ins for SDL Trados Studio, Kilgray memoQ, and Memsource.


LetsMT! Machine Translation Provider,En_Lv TM - Translation Results
Project Settings...
1
2
3
Let's MT
LetsMT! Machine Translation Provider,En_...
LetsMT! Machine Translation Provider,En_...
Comments
Demo.docx.sdlxliff [Translation]*
Demo.docx
P

Click another button to see the demo again.

1	Click the button to see the demo.	82%	Noklikšķiniet uz pogas, lai skatītu demonstrāciju.
2	Click another button to see the demo again.	AT	Noklikšķiniet uz citas pogas, lai skatītu demonstrāciju.

LetsMT! Machine Translation Provider 2011.04.29. 12:26:11

1	Click the button to see the demo.	AT	Noklikšķiniet uz pogas, lai skatītu demonstrāciju.	P
2	Click another button to see the demo again.			P
3	Click the button to see the demo again.			P

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

Tilde MT can also be used in a true post-editing scenario where the whole document or project is machine translated and presented to the human for post-editing. The same CAT tool plug-in can be used to pre-translate the whole text.

Another option to pre-translate the whole document is the file translation. Users can upload files on the Tilde MT web page[2] and get them translated with all the original formatting preserved. Currently Tilde MT supports TMX, XLIFF, TTX, and DOCX file formats.

5. MT impact on productivity

This section of the paper summarizes the results of our productivity evaluation experiments in localization tasks (Skadiņš et al., 2014). The aim of our experiments was to assess the impact of MT on translators’ productivity and translation quality in a typical localization scenario. In order for the MT application to be useful it has to allow achieving significant productivity improvement in the translation process, that is, decreasing the total time spent on translation while keeping the required level of quality. To assess this we measure:

- translators’ productivity,
- quality of translation,
- time spent on identification and correction of errors in the translations.

Unlike in many other post-editing experiments (e.g., Plitt and Masselot, 2010; Teixeira, 2011) where automatic tools were used to measure time spent on individual activities, to log translator key strokes, etc., we evaluated productivity and quality in a realistic working environment. We applied the typical everyday translation workflow using the same tools for process management, time reporting, and quality checking as in everyday work.

We performed experiments in two scenarios:

- *Scenario 1* – translation using TM only (the baseline scenario).

[2] http://tilde.com/mt

- *Scenario 2* – translation using TM and MT; MT suggestions are provided for every translation unit that does not have a 100% match in TM.

5.1. Data for evaluation

Evaluation was performed in the software localization domain for translations from English into target language(s). The following criteria were applied in selecting the source text (documents) for evaluation: (1) the documents have not been translated before, (2) about 50% of the documents contain at least 95% new words, (3) about 50% of documents contain sentences with different level of fuzzy matches, and (4) the size of each document has to be about 1,000 weighted words on average.

All documents were split into 2 equally sized parts to perform two translation scenarios described above. Texts were selected from user assistance and user interface sub-domains. In the first experiment the following requirements were applied for the selection of the test set: (1) only plain text documents containing no formatting tags, (2) documents related to the topics of the data on which the SMT systems are trained, and (3) documents with a similar style and terminology as in the training data used for generating SMT. For the second experiment a different test set was selected: (1) documents containing text with mark-up (formatting or tags, placeholders, etc.), (2) documents have to be in the same domain as the data on which the SMT systems were trained, but sub-domains may differ, and (3) documents that have different style and terminology to the training data.

The different approaches in the selection of the test sets make the two experiments not comparable. But that was to be expected, as the goals of the two experiments differ significantly.

5.2. Evaluation Process

The evaluation process was the same for all languages. At least 5 translators were involved with different levels of experience and average (or above average) productivity. All translators were trained to use MT systems and SDL Trados Studio 2009 or 2011 in their translation work before the evaluation process started. The LetsMT plug-in for the SDL Trados 2009 (or 2011) CAT environment was used in all experiments.

In both scenarios, translators were allowed to use whatever external resources they needed (dictionaries, online reference tools, etc.), just as during regular operations. Translators performed the test without interruption and without switching to other translation tasks during their working day – 8 hours – because splitting the time into short periods would not show reliable evaluation results. Each scenario was performed on a different working day. The time spent for translation was manually reported. To avoid any “start-up” impact, in *Scenario 2* we removed from the result analysis the first translation task performed by each translator.

5.3. Productivity and Quality Assessment

Each translator’s productivity was calculated as a number of weighted words translated per hour. The translation quality for each document was evaluated by at least 2 experienced editors. Editors were not aware of the scenario used (whether MT was applied or not).

The quality of translation is measured by filling in a QA form in accordance with the QA methodology based on the LISA (2008) QA model. The evaluation process involves inspection of translations and classifying errors according to the error categories.

The productivity and quality of work was measured and compared for every individual translator. An error score was calculated for every translation task 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 1,000 weighted words as follows:

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

where n is a weighted word count in a translated text, e_i is a number of errors of type i and 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 the type *comprehensibility* (an error that obstructs the user from understanding the information; very clumsy expressions) have the weight 3, while errors of the type *omissions/unnecessary additions* have the weight 2. Depending on the error score the translation is assigned a translation quality grade – superior, good, mediocre, poor, or very poor.

5.4. Experiment 1

The goal of the first experiment was to test the hypothesis that MT can be beneficial in everyday operations of translators and that it can increase their productivity. The experiment was performed for the English-Latvian language pair with a domain specific SMT system. We used the best available MT system (Skadiņš et al., 2010) trained on both domain specific training data and out-of-domain publicly available training data, such as, DGT-TM and OPUS EMEA (Tiedemann, 2009). The system also includes knowledge about Latvian morphology. For training and running SMT systems we used the cloud-based platform LetsMT.

For automatic MT system evaluation we used the BLEU metric. The IT domain tuning (2,000 sentences) and testing (1,000 sentences) data were automatically filtered out from the training data before the training process. The BLEU score of the MT system was 70.37.

The MT system used in the evaluation was trained using specific vendor translation memories[3] as a significant source of parallel corpora. Therefore, the SMT system may be considered slightly biased to a specific IT vendor, or a vendor specific narrow IT domain. The evaluation set contained in equal proportions texts from this vendor and a different vendor whose translation memories were not included in the training data.

The results are assessed by analyzing average values of translators' productivity and the error score for translated texts. Usage of MT suggestions in addition to the use of TMs increased the productivity of translators in all evaluation experiments. The average productivity in *Scenario 1* was 550 weighted words per hour, and 731 weighted word per hour in *Scenario 2*. This means that an average productivity increase of 32.9% was observed. There were significant productivity differences in the various translation tasks. The standard deviations of productivity in the baseline and MT scenarios were 213.8 and 315.5, respectively. Significant differences in the results of different translators have been observed; the results vary from a 64% increase in productivity to a 5% decrease in productivity for one of the translators. Further analysis is necessary, but most likely the differences are caused by the working patterns and skills of individual translators.

At the same time, the error score increased from 20.2 to 28.6 still remaining at the quality grade "good". We have not performed a detailed analysis of the reasons causing an increase in the error score, but this can be explained by the fact that translators tend to trust suggestions

[3] 1.29 M sentences of in-domain data.

coming from the CAT tool and do not sufficiently check them, even if they are marked as a MT suggestions.

5.5. Experiment 2

Although our first experiment showed significant productivity increase, translators were reluctant to use MT in their everyday work. The reason was due to various mark-ups (tags, placeholders, etc.) which are very frequent in real-life translation segments, but were not properly handled by the MT system, thereby requiring a lot of additional post-editing efforts.

The goal of the second experiment was to evaluate a more complex translation scenario where source documents contain formatting tags, placeholders and differ in the used terminology and language style, and thus are slightly out-of-domain for the SMT system than in the previous experiment. We performed this experiment to analyze the LetsMT platform and SMT systems trained on it in a difficult scenario, to find more detailed beneficial aspects of MT usage in localization workflows and to identify areas that require improvements. The experiment was performed for three language pairs: English-Estonian, English-Latvian and English-Lithuanian.

All three MT systems were trained on proprietary parallel corpora in the IT domain (consisting of user manuals, user interface strings, technical documents, etc.). See Table 3 for the size of the parallel corpora for translation model training. Two different English-Latvian MT systems were trained; the second MT system (v2) had much better support for different formatting tags, URLs, numbers and other non-translatable units. The results of the SMT system automatic evaluation are given in Table 3.

MT System	Size (sentences)	BLEU score	METEOR score
EN-LV (v1)	1.70 M	69.57	0.48
EN-LV (v2)	3.80 M	66.98	0.46
EN-LT	2.14 M	59.72	0.43
EN-ET	3.56 M	55.88	0.40

Table 3. Results of automatic MT system quality evaluation for the second experiment.

We created the evaluation data sets by selecting documents in the IT domain that had not been translated by the translators before the evaluation. Similarly to the first experiment, this ensured that translation memories did not contain the translatable segments. We also selected documents aiming at different target audiences (system administrators, programmers, everyday users) as well as from vendors contrasting to the ones whose translation memories were used in the training of SMT systems (usually having different translation guidelines and writing styles). This ensured that the selected texts were of different linguistic characteristics (including syntax, terminology usage, style, etc.), thus making the translation task more difficult for the SMT systems. Documents for translation were selected if they contained c.a. 1,000 weighted words each and had formatting tags (on average in ¼ to ⅓ of all translation segments).

Following the evaluation procedure of the first experiment, we analyzed the average values for productivity and the error score for translated texts.

Language pair	Productivity changes	Standard deviation changes in %
EN-LV (v1)	-3.10% ± 5.76%	20.80%
EN-ET	-4.70% ± 7.53%	27.17%
EN-LT	-3.76% ± 8.11%	29.28%

Table 4. Productivity changes from *Scenario 1* to *Scenario 2* with a 95% confidence interval

Bearing in mind the complexity of this experiment (formatting tags, more complex language and slight subdomain deviations from the data the SMT system is trained on), the results suggest that the average productivity slightly decreases for all language pairs; however, this cannot be statistically proved in a 95% confidence interval (as shown in Table 4). The large confidence interval is caused by the significant productivity differences (as shown by the changes of the standard deviation of productivity) in the various translation tasks.

The quality review results for all three language pairs are given in Table 5. The results show a minor decrease of translation quality, from 18.7 to 23.0 points for English-Latvian and from 17.0 to 22.7 points for English-Lithuanian. For English-Estonian the quality of translated texts slightly increased (from 12.9 to 12.0), which is mainly because of "superior" quality ratings for two translators. Although for two language pairs we see a slight drop, the quality evaluation grade is still at the level "good", which is acceptable for production.

Language pair	**Error score Scenario 1**	**Error score Scenario 2**
EN-LV (v1)	18.7	23.0
EN-LT	17.0	22.7
EN-ET	12.9	12.0

Table 5. Linguistic quality evaluation results of the second experiment

After evaluation, translators submitted in-formal feedback describing their SMT post-editing experience. Three main directions for further improvements were evident:

- In many cases segments with formatting tags were not translated correctly due to limitations and errors in our implementation of the tag translation functionality.
- As every segment was sent to the MT platform only at the time of its translation, translators had to wait up to 3 seconds while an SMT translation suggestion was provided. Pre-translation or increase of MT speed would solve this problem.
- SMT made a lot of errors in handling and translating named entities, terminology, numbers, non-translatable phrases (e.g., URLs, file paths, etc.).

Since the second experiment, we have actively worked to address the issues raised by the translators. Bugs in the tag translation framework have been fixed, specific non-translatable named entity (e.g., directory paths, URLs, number sequences, etc.) as well as some structured named entity (e.g., dates, currencies) handling has been implemented in the LetsMT platform, and most importantly SMT pre-translation was enabled for the translators. Our preliminary analysis on a small-scale evaluation scenario (following the guidelines of the second experiment) for English-Latvian with two involved translators and 16 translation tasks (8 translation tasks per scenario) shows that the average productivity using the improved LetsMT platform increases from 16.7% up to 35.0% when using SMT support over manual translation without SMT support. This suggests that even for very difficult scenarios SMT systems can be beneficial and lead to significant productivity increases.

6. Conclusions

The results of our experiments 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 experiments on terminology integration in SMT system training showed a quality improvement of up to 24% over the baseline system. However, considering that re-training of

MT engines may not be economically justifiable in many situations, dynamic integration methods that can re-use pre-trained SMT systems and integrate terminology *on-the-fly* are necessary. This is where we focus our current (TaaS, 2014) and future research efforts on improving methods for terminology integration in SMT systems.

The evaluation shows that the use of the SMT suggestions in addition to the translation memories in the SDL Trados CAT tool leads to the increase of translation performance by up to 32.9% while maintaining an acceptable quality of 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. At the same time, usability of MT is significantly lower for texts with rich formatting and in-line tags.

Error rate analysis shows that overall usage of MT suggestions decreases 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 leading to these results has received funding from the research project “2.6. Multilingual Machine Translation” of EU Structural funds, contract nr. L-KC-11-0003 signed between ICT Competence Centre and Investment and Development Agency of Latvia.

References

Babych, B., & Hartley, A. (2003). Improving Machine Translation Quality With Automatic Named Entity Recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*.

Bertoldi, N., Haddow, B., & Fouet, J.-B. (2009). Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, 91(1), 7–16.

Bouamor, D., Semmar, N., & Zweigenbaum, P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 674–679).

Carl, M., & Langlais, P. (2002). An Intelligent Terminology Database as a Pre-processor for Statistical Machine Translation. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14* (pp. 1–7).

Deksne, D., & Skadiņš, R. (2012). Data Pre-Processing to Train a Better Lithuanian-English MT System. In A. Tavast, K. Muischnek, & M. Koit (Eds.), *Frontiers in Artificial Intelligence and Applications, Volume 247: Human Language Technologies – The Baltic Perspective* (pp. 36–41). IOS Press. doi:10.3233/978-1-61499-133-5-36

Denkowski, M, & Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 138–145). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hálek, O., Rosa, R., Tamchyna, A., & Bojar, O. (2011). Named Entities from Wikipedia for Machine Translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies* (ITAT 2011) (pp. 23–30).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177–180). Stroudsburg, PA, USA: Association for Computational Linguistics.

Koehn, P., & Schroeder, J. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 224–227). Prague, Czech Republic.

Koehn, P., Birch, A., & Steinberger, R. (2009). 462 Machine Translation Systems for Europe, *Proceedings of MT Summit XII.*

Lewis, W. D., Wendt, C., & Bullock, D. (2010). Achieving Domain Specificity in SMT without Overt Siloing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 2878–2883).

LISA. (2008). LISA QA model. Retrieved from http://web.archive.org/web/20080124014404/http://www.lisa.org/products/qamodel/

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

Pinnis, M., Gornostay, T., Skadiņš, R., & Vasiļjevs, A. (2013a). Online Platform for Extracting, Managing, and Utilising Multilingual Terminology. In *Proceedings of the Third Biennial Conference on Electronic Lexicography*, eLex 2013 (pp. 122–131). Tallinn, Estonia: Trojina, Institute for Applied Slovene Studies (Ljubljana, Slovenia) / Eesti Keele Instituut (Tallinn, Estonia).

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., & Gornostay, T. (2012). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering* (TKE 2012) (pp. 193–208). Madrid.

Pinnis, M., Skadiņa, I., & Vasiļjevs, A. (2013b). Domain Adaptation in Statistical Machine Translation Using Comparable Corpora: Case Study for English Latvian IT Localisation. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2013) (pp. 224–235). Samos, Greece: Springer Berlin Heidelberg.

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

Rehm, G. & Uszkoreit, H., editors. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*: Springer, Heidelberg etc. 32 volumes on 31 European languages.

Skadiņš, R., Goba, K., & Šics, V. (2010). Improving SMT for Baltic Languages with Factored Models. In *Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications*, Vol. 2192 (pp. 125–132). Riga: IOS Press.

Skadiņš, R., Pinnis, M., Vasiļjevs, A., Skadiņa, I., & Hudik, T. (2014). Application of Machine Translation in Localization into Low-Resourced Languages. In M. Tadić, P. Koehn, J. Roturier, & A. Way (Eds.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation EAMT2014* (pp. 209–216). Dubrovnik: European Association for Machine Translation. Retrieved from http://hnk.ffzg.hr/eamt2014/EAMT2014_proceedings.pdf

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas* (pp. 223–231). Cambridge, MA, USA.

Steinberger, R, Eisele, A, Klocek, S, Pilos, S, Schlüter, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. *Proceedings of the 8th international conference on Language Resources and Evaluation* (LREC'2012), Istanbul.

TaaS. (2014). Public Deliverable D4.4 Integration with SMT Systems. *TaaS Project: Terminology as a Service*.

Teixeira, C. (2011). Knowledge of provenance and its effects on translation performance in an integrated TM/MT environment. *Proceedings of the 8th International NLPCS Workshop - Special theme: Human-Machine Interaction in Translation* (pp. 107-118).

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing (vol V)* (pp. 237-248), Amsterdam/Philadelphia: John Benjamins.

TTC. (2013). Public Deliverable D7.3: Evaluation of the Impact of TTC on Statistical MT (p. 38). *TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora*.

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: Cloud-Based Platform for Do-It-Yourself Machine Translation. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 43–48). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P12-3008