Dynamic Terminology Integration Methods in Statistical Machine Translation

Mārcis Pinnis

Tilde, Vienibas gatve 75a, Riga, Latvia
University of Latvia, 19 Raina Blvd., Riga, Latvia
`marcis.pinnis@tilde.lv`

Abstract

In this paper the author presents methods for dynamic terminology integration in statistical machine translation systems using a source text pre-processing workflow. The workflow consists of exchangeable components for term identification, inflected form generation for terms, and term translation candidate ranking. Automatic evaluation for three language pairs shows a translation quality improvement from 0.9 to 3.41 BLEU points over the baseline. Manual evaluation for seven language pairs confirms the positive results; the proportion of correctly translated terms increases from 1.6% to 52.6% over the baseline.

1 Introduction

In professional translation services, correct and consistent handling of terminology is an important indicator of translation quality. However, pure statistical machine translation (SMT) systems, such as, Moses (Koehn et al., 2007) in a general scenario cannot ensure correct and consistent handling of terminology, because statistics of large amounts of data are difficult to control if not constrained by means of, e.g., bilingual term collections or translation model or language model adaptation techniques. In cases where the context is too ambiguous (e.g., if an SMT system receives just a short translation segment or the SMT system's models are limited in the possibilities to analyse larger context) or when external knowledge is re-quired, it can be impossible for an SMT system to guess the correct translation.

In the localisation industry customers often provide their own term collections that have to be strictly used during translation to ensure correct and consistent usage of terminology. Obviously, such collections may contain term translations that are rated as unlikely (in certain contexts) by an SMT system's models or they may even be missing in the models at all if custom adaptation of the models using the customers' provided data is not performed. If such SMT systems would be integrated in localisation service workflows, it would not be possible to ensure high terminology translation quality in the SMT suggestions. Therefore, effective methods that can benefit from custom term collections are necessary.

Researchers have tried to address the terminology integration challenge directly by using in-domain term collections and indirectly by tackling the broader challenge of domain adaptation. Significant research efforts have been focussed on using in-domain parallel and monolingual corpora (that contain in-domain terminology) to perform SMT system translation and language model adaptation to specific domains (to name but a few, Koehn & Schroeder (2007), Bertoldi & Federico (2009), Hildebrand et al. (2005), and many others). Terminology integration has been also indirectly addressed by research on multi-word unit integration in SMT. E.g., Bouamor et al. (2012) showed that for French-English it is enough to simply add multi-word unit pairs to the parallel corpus; however, they observed a limited gain of +0.3 BLEU (Papineni et al., 2002) points. In terms of direct terminology integration, Pinnis & Skadiņš (2012) have shown that the addition of terms to the parallel corpus and the introduction of a bilingual termi-

nology identifying feature in the translation model can significantly improve translation quality of an out-of-domain system (up to +2.13 BLEU points). Their method specifically addressed morphologically rich languages by identifying terms in different inflected forms using stemming tools. Similar work that shows significant quality improvements has been recently performed by Arcan et al. (2014a) for the English-Italian language pair. They use a term collection to create a "fill-up" translation model that consists of a pre-trained SMT system's phrase table merged with a phrase table created from the bilingual terminology. However, all these methods require to re-train the whole SMT system (or at least re-tune the SMT system) if new in-domain data becomes available. For many translation tasks such a scenario is not economically justifiable. Furthermore, if we have already trained a relatively good SMT system (let it be a general domain system or a close-domain system to the domain that is needed), we should be able to tailor it to the required domain with the help of just the right bilingual terminology.

Consequently, considerable research efforts have been focussed also on dynamic integration methods for term collections in SMT that do not require re-training of SMT systems. For instance, the *Moses* SMT system supports input data (in the Moses XML format) that is enriched with externally generated translation candidates. Using this methodology, Carl & Langlais (2002) used term dictionaries to pre-process source text and achieved an increase in translation quality for the English-French language pair. Similarly, Arcan et al. (2014a) identify exactly matched terms and provide translation equivalents from the Wiki Machine[1] by performing context-based disambiguation if there are multiple translation equivalents for a single term for English-Italian. Babych & Hartley (2003) showed that inclusion of certain named entities in "*do-not-translate*" lists allowed to increase translation quality for the English-Russian language pair. Recently dynamic translation and language models (Bertoldi, 2014) have been investigated for integration of terminology into SMT (Arcan et al., 2014b) for English-Italian. It is evident that most of the related research has, however, mostly focused on languages with simple morphology or translation of phrases that are rarely

translated or even left untranslated. A study in the FP7 project TTC (2013) showed that for English-Latvian such simplified methods do not yield positive results. Hálek et al. (2011) came to the same conclusion in their work on English-Czech named entity translation. This means that for morphologically rich languages more linguistically rich methods are necessary.

In this paper, the author proposes a workflow for dynamic terminology integration in SMT systems that allows to: 1) identify terms in source text (i.e., translation segments or even large documents with Moses XML tags) that is sent to the SMT system for translation, 2) generate inflected forms of terms using corpus-based and morphological synthesis-based methods, and 3) rank term translation candidates. The methods proposed have been evaluated in two different scenarios using automated SMT quality metrics for three language pairs and by performing manual comparative evaluation for seven language pairs (from English into Estonian, French, German, Italian, Latvian, Lithuanian, and Spanish). The results will show that the proposed methods are able to improve terminology translation quality and the overall sentence translation quality for morphologically rich languages. For evaluation purposes, the author uses the LetsMT SMT platform (Vasiļjevs et al., 2012), which is based on the Moses SMT system.

The paper is further structured as follows: section 2 describes the dynamic terminology integration workflow and the different modules for source text pre-processing, section 3 describes our automatic and manual evaluation efforts, and section 4 concludes the paper.

2 Dynamic Terminology Integration Workflow

The idea of the dynamic terminology integration scenario (conceptually depicted in Figure 1) is that users (e.g., translators when using SMT capabilities in a computer-assisted translation (CAT) environment, Web site owners when integrating SMT widgets in their Web sites, etc.) have to be able to assign custom bilingual term collections to pre-trained SMT systems of the LetsMT platform when there is a need to translate some content. To ensure this functionality the author utilises the capability of the Moses decoder to translate input data in the Moses XML format and introduce a new source text pre-processing workflow before

---

[1]The Wiki Machine is available online at: https://bitbucket.org/fbk/thewikimachine
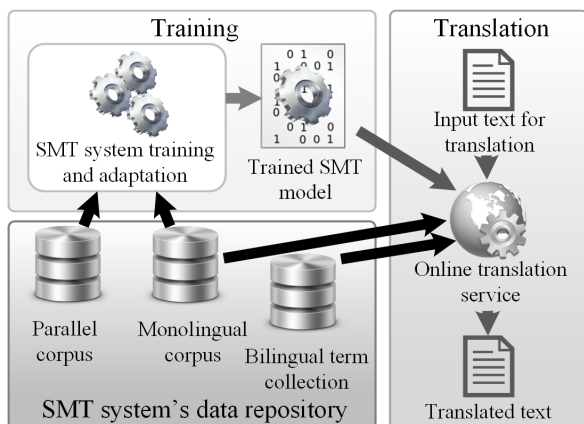
Figure 1: The conceptual design of dynamic terminology integration in SMT systems

decoding the content with the Moses decoder. The workflow (depicted in Figure 2) consists of three exchangeable modules that 1) use a bilingual term collection provided by the user to identify terms in the source text using term identification methods (see section 2.1), 2) generate inflected forms of the translations of the identified terms (see section 2.2), and 3) assign translation confidence scores to translation candidates and enrich the source text with the generated translation candidates (see section 2.3). After pre-processing the terminology enriched content is translated with the Moses decoder by explicitly using the provided translation candidates.
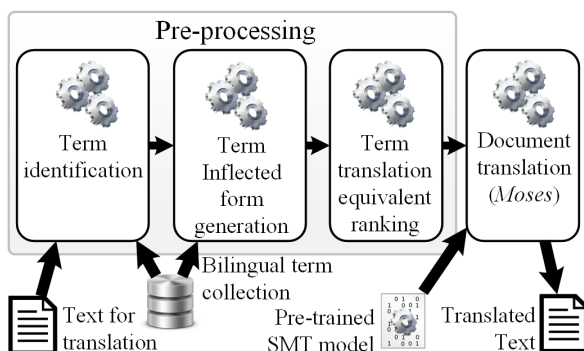

Figure 2: Source text pre-processing workflow

When using SMT capabilities in on-line scenarios, let it be translation of full Web pages, on-line pre-translation of following translation segments in CAT tools (e.g., the MateCat platform by Federico et al. (2014)), or any other scenario that requires quick SMT response, an important factor to be considered for dynamic integration methods is their impact on the overall speed of translation.

For successful SMT integration in localisation scenarios, it is crucial that SMT systems can provide translations quickly as translator performance will decrease if the translators will have to wait for SMT suggestions (Skadiņš et al., 2014). To ensure that the effect on the overall translation speed is minimal, compromises between how linguistically rich term identification and ranking has to be or whether or not to perform the inflected form generation for terms in an off-line mode (i.e., when uploading a term collection to the SMT platform) have to be met. The proposed workflow allows to decide whether processing speed or linguistic richness is of greater importance.

2.1 Term Identification

The first task that has to be performed when pre-processing source text using a bilingual term collection is to identify terms. For this purpose, three methods were investigated:

- The first method (*TWSC*) performs term identification using the linguistically and statistically motivated term extraction tool *TWSC* (Pinnis et al., 2012). TWSC 1) morpho-syntactically tags and lemmatises the source text, 2) extracts term phrases that match morpho-syntactic term phrase patterns (most commonly, noun phrases), 3) performs statistic ranking using co-occurrence measures and reference corpora statistics, and 4) tags terms in a document by prioritising longer phrases. Then, the extracted term phrases are looked-up in the term collection by comparing their lemma and part of speech sequences. If the terms in the term collection do not contain morpho-syntactic information, terms are morphologically analysed and lemmatised, after which all matching term phrase patterns from TWSC are identified and used for look-up purposes.

- As the source text may be too short to perform statistical analysis and because we only search for term phrases that are included in the user provided term collections, the second method (*Valid Phrase-Based Term Identification* or *Phrase*) starts by performing the two steps from TWSC, however then it directly looks-up, whether the morpho-syntactically valid term phrases actually correspond to a term from the term collection.

91

- The first two methods rely heavily on linguistic tools that can significantly affect the translation speed. Therefore, the third method (*Fast Term Identification* or *Fast*) performs a left-to-right search in the source text using minimal linguistic support from language-specific stemming tools to identify terms in different inflected forms.

2.2 Inflected Form Generation

The next pre-processing step after term identification is the generation of translation candidates for the identified terms. Previous research (Nikoulina et al., 2012; Carl & Langlais, 2002; Babych & Hartley, 2003) on source text pre-processing has not given special attention to this question, because the bilingual term collections already "*provide*" translation equivalents. However, the issue is that the terms that are provided in the bilingual term collections are usually in their canonical forms. For morphologically rich languages the canonical forms in many contexts are not the required inflected forms. Because of the focus on language pairs that do not require (or require very limited) morphological generation (e.g., English-French, English-German, etc.), previous research has not seen the need to address these issues. Therefore, the author investigated three different methods for acquisition of inflected forms of terms:

- The first method (*Synthesis*) uses a morphological analyser and synthesiser and inflected form generation rules to generate inflected forms of a term from its canonical form. E.g., the Latvian term 'datu tips' (in English: 'data type') corresponds to the term phrase pattern '`^N...g.*  ^N.*`' consisting of two nouns (the first word is in a genitive case). The term phrase pattern corresponds to the inflection rule '`*****  ***00`'. The rule specifies that the first word has to be kept as is (the '*'), however the second word is allowed to be in any inflected form of a noun ('0' indicates that any value for a morphological category is acceptable; in the positional tagset used for Latvian the fourth and fifth positions correspond to case and number). The rules allow defining also morpho-syntactic agreements between different morphological categories (e.g., in Latvian adjectives in a noun phrase have to have the same gender, number, and case as the head noun). For Latvian there

are in total 18 inflection rules specified for 99 term phrase patterns from TWSC.

- The second method (*Corpus*) is language independent and relies on the SMT system's monolingual corpus (e.g., the corpus that is used for language modelling) to identify inflected forms of terms using a similar method to the *Fast Term Identification*.

- Both previous methods may not be able to generate inflected forms for all terms. For instance, the first method may lack a term phrase pattern necessary for a specific term, whereas, when applying the second method, some inflected forms may be missing in the corpus or the stemming tool may not be able to identify all forms. Therefore, the third method (*Combined*) is a combination (using union) of both previous methods.

2.3 Term Translation Equivalent Ranking

As the last pre-processing step, the generated translation candidates have to be ranked by assigning translation confidence scores. For this purpose two methods were investigated:

- The first method (*Equal*) assigns equal translation likelihood scores to all translation candidates of a term. This method is used as a baseline method for translation candidate ranking. When assigning equal weights to all translation candidates, we rely on the language model to select the most likely translation.

- The second method (*Simple*) uses a large monolingual corpus and calculates for each translation candidate of a term its relative frequency among all translation candidates of the term. This method allows assigning higher scores for more common translations.

It is evident that both methods rely only on the language model and important statistics that may come from the translation model (e.g., source to target language transfer information) are lost. We also lose important information from the source language's context as that could help identifying, which translation candidate is more likely in a given context. However, the potentially more sophisticated methods are left for future work.

3 Evaluation

To evaluate the dynamic terminology integration methods, two evaluation tasks were carried out: 1) automatic evaluation that identifies the combination of the different methods that allows achieving the highest results, and 2) manual evaluation that focusses on term translation qualitative analysis using production SMT systems and an authoritative term collection. The following subsections describe both evaluation efforts.

3.1 Automatic Evaluation

The automatic evaluation was performed for three language pairs (English-German, Latvian, and Lithuanian) using general domain SMT systems that were trained in the LetsMT platform using the DGT-TM parallel corpus (Steinberger et al., 2012) (the releases of 2007, 2011, and 2012). For evaluation, the author uses a proprietary parallel corpus of 872 sentence pairs in the automotive domain (technical documentation from car service manuals). The original data set was available for English-Latvian, therefore, the remaining two data sets for German and Lithuanian were prepared by professional translators. For English-Latvian an in-domain tuning set of 1,745 sentence pairs was available; for the remaining systems held-out sets of 2,000 sentence pairs from the training data were used for SMT system tuning. The results of the baseline systems are given in Table 1. It is evident that the results for English-Latvian are significantly higher (although still relatively low) than for the other language pairs. This is mainly due to the fact that an automotive domain tuning set was available for the English-Latvian experiments. As the results for the other language pairs are very low, Table 1 includes also automatic evaluation results using 1000 held-out sentence pairs from the DGT-TM corpus to show that the systems on in-domain data perform relatively well. This shows just how different the writing styles and the language complexity between different domains can be.

Lang. pair	EN-DE	EN-LT	EN-LV
BLEU (a)	8.27	6.94	12.68
BLEU (g)	54.03	48.12	-

Table 1: Baseline system performance ("*(a)*" - automotive domain evaluation sets; "*(g)*" - SMT system in-domain evaluation sets from the DGT-TM corpus)

Next, the author analysed, which pre-processing configuration allows achieving better results (see Figure 3). This analysis was performed for English-Latvian using a term collection that was created by a professional translator from the tuning-data. The term collection consists of 644 term pairs (terms were included only in their canonical forms). The results show that all combi-
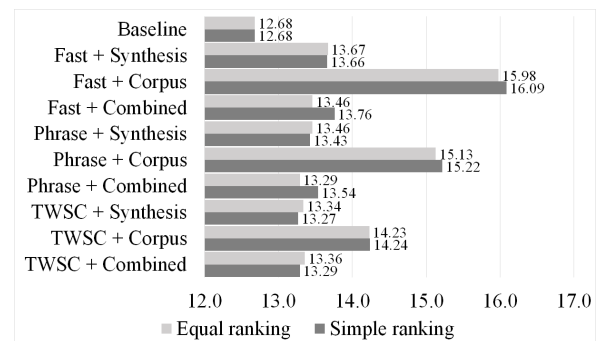

Figure 3: Automatic evaluation results using three different term collections for English-Latvian (BLEU scores)

nations performed better than the baseline system. It is evident that the *Fast Term Identification* allows achieving better results than the other term identification methods. The method also allows to identify more terms in the source text (1,404; compared to 1,261 for *Phrase* and 620 for *TWSC*). We see also that the *Synthesis* method for inflected form generation achieves lower results than the *Corpus* method for which there are two possible reasons: 1) data ambiguity for the SMT system by providing significantly more inflected forms is increased, and 2) the implemented ranking methods do not allow effectively estimating, which inflected form is more or less likely due to not taking the language transfer characteristics into account.

Next, professional translators were asked to prepare professional term collections for English-German (692 term pairs) and English-Lithuanian (662 term pairs) and performed automatic evaluation experiments. The results in Figure 4 are limited to the configurations with '*Corpus+Simple*' that showed to achieve the best results for English-Latvian.

3.2 Manual Evaluation

The automatic evaluation showed positive results. However, the SMT systems in the baseline scenario achieved relatively low scores and the term collections were relatively small (although fo-
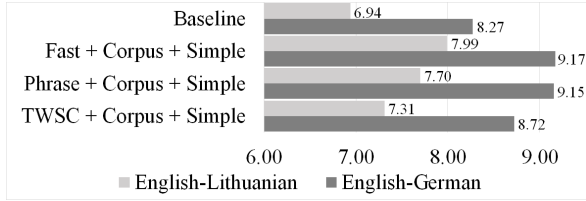
Figure 4: Automatic evaluation results using different term identification methods and corpus-based inflected form generation and ranking

cussed to a narrow domain). Therefore, the manual evaluation was performed for seven language pairs using production level in-domain SMT systems (contrary to out-of-domain systems before) in the information technology domain. For terminology integration, the freely available Microsoft Terminology Collection[2] was used.

As the term collection contains many ambiguous terms that can be confused with general language words and phrases (e.g., '*AND*', '*about*', '*name*', '*form*', '*order*', etc.), it is important to filter such candidates out as the dynamic integration workflow (contrary to methods that perform SMT system model adaptation) is sensitive to the level of ambiguity of the included terms. The collections for the different language pairs were filtered using a term pair specificity estimation method that is based on inverse document frequency (IDF) scores (Spärck Jones, 1972) from a broad domain corpus. The formula is given in (1); it was first introduced by Pinnis & Skadiņš (2012).

$$R(p_s, p_t) = \\ min\left(\sum_{i=1}^{|p_s|} IDF_s(p_s(i)), \sum_{j=1}^{|p_t|} IDF_t(p_t(j))\right) \quad (1)$$

The baseline system performance and the term collection statistics are given in Table 2.

Lang. pair	BLEU	Terms (filtered)	Terms (initial)
EN-ES	74.61	18,871	23,094
EN-FR	68.76	19,665	24,160
EN-ET	55.23	10,175	12,648
EN-LT	60.42	10,352	12,726
EN-LV	66.98	10,497	12,926
EN-RU	60.79	18,416	22,669
EN-DE	61.35	20,308	24,997

Table 2: Baseline system performance (on 1,000 held-out sentence pairs) and statistics of the term collections before and after filtering

---

[2]The Microsoft Terminology Collection can be downloaded from: http://www.microsoft.com/Language/en-US/Terminology.aspx

The manual evaluation is performed by comparing the SMT system performance without (the baseline scenario) and with (the improved scenario) integrated terminology. The '*Fast+Corpus+Simple*' configuration was used in this experiment. The evaluation data for each language pair consists of 100 in-domain sentences for which the outputs of the SMT systems in the two scenarios differed (different translations were produced in average for 56% of sentences). For each language pair two professional translators were involved in the evaluation.

For the evaluation, translators were asked to perform three ratings:

- For each sentence, translators had to decide which scenario produced a better translation. If both scenarios produced translations of equal quality, the translators had to decide whether both scenarios produced acceptable or not acceptable translations.

- Similarly to the sentence level, for each term that was identified in the source text using the '*Fast*' method, translators had to decide which scenario produced a better translation.

- The first two are quantitative analysis measures, therefore as a third rating translators were asked to rate the term translation quality in both scenarios separately. The translators had to decide whether the term is translated correctly, whether a wrong inflectional form is used, whether it is not translated, whether it is split up or its words are in a wrong order, whether a wrong lexical choice is made, whether the marked phrase is actually not a term and has been wrongly identified as a term, or whether there is another issue.

The sentence level evaluation summary in Table 3 shows that the translations of the improved scenario were preferred more for six language pairs. Because of spatial restrictions, the paper features only results from the analysis where evaluators were in full agreement. It is evident that the task of comparing sentence level quality is a very challenging task for evaluators, because the Free Kappa (Randolph, 2005) agreement scores are mainly in the levels of fair to moderate.

The term level evaluation summary is given in Table 4. It is evident that translation quality has improved over the baseline scenario for all language pairs evaluated. Even more, the agreement

Lang. pair	Bas.	Imp.	Both	None	Total	Free Kappa
EN-ES	11	8	15	19	53	0.38
EN-FR	8	21	35	18	82	0.16
EN-ET	8	16	3	36	63	0.50
EN-LT	6	8	23	16	53	0.37
EN-LV	1	9	9	57	76	0.68
EN-RU	9	17	7	27	60	0.47
EN-DE	5	15	29	9	58	0.45

Table 3: Evaluation summary for sentence level ratings where evaluators were in agreement

Lang. pair	Bas.	Imp.	Both	None	Total	Free Kappa
EN-ES	4	34	77	0	115	0.64
EN-FR	4	71	141	4	220	0.44
EN-ET	21	51	53	0	125	0.70
EN-LT	1	40	54	3	98	0.49
EN-LV	6	46	67	4	123	0.75
EN-RU	1	49	93	0	143	0.82
EN-DE	2	30	87	0	119	0.70

Table 4: Evaluation summary for term level ratings where evaluators were in agreement

scores for evaluators show that the task of comparing in which system terms were translated better was fairly easy and in general well understood.

The summary of the term translation quality evaluation for the individual scenarios is given in Table 5. The results show that the proportion of correct term translations has improved for all language pairs from +1.6% for English-Estonian to +52.6% for English-Lithuanian. The minimal improvement for English-Estonian is mainly due to selection of wrong inflected forms (which is a lesser quality issue, but an issue nonetheless) rather than wrong term lexical choices (which is a greater quality issue). The author believes that the relatively low performance for English-Estonian is caused by the under-performance of the word stemming component for Estonian that is used for inflectional form acquisition for terms (however, deeper investigation is necessary). It is evident that in terms of using the correct lexical choice, the quality has improved from +26.4% for English-German to +65.2% for English-Lithuanian. This means that the method allows ensuring terminology translation consistency better than in the baseline scenario.

4 Conclusions

The paper presented a source text pre-processing workflow for dynamic terminology integration in SMT systems. To evaluate the methods, the au-

thor performed automatic evaluation in the automotive domain. The results show that the best combination of pre-processing methods achieved a translation quality improvement from 0.9 to 3.41 BLEU points (depending on the language pair) over the baseline scenario. Manual evaluation for seven language pairs indicates that the proportion of correctly translated terms increased from 1.6% to 52.6% over the baseline scenario.

Although the results are positive, the best results were achieved using lightly linguistic methods (i.e., stemming tools). The linguistically more advanced methods could either identify less terms or produced too many inflected forms of terms, thus making it more difficult for the SMT decoder to select the correct form. The author believes that a language transfer based term ranking method and a method that combines the different term identification methods could improve the results even further. However, this is an area for future work.

5 Acknowledgements

References

Arcan, M., Giuliano, C., Turchi, M., and Buitelaar, P. 2014a. Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation. In *Proceedings of CompuTerm 2014*.

Arcan, M., Turchi, M., Tonelli, S., and Buitelaar, P. 2014b. Enhancing Statistical Machine Translation with Bilingual Terminology in a CAT Environment. In *Proceedings of AMTA 2014* (pp. 54–68).

Babych, B., and 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. 2014. Dynamic models in Moses for Online Adaptation. In *The Prague Bulletin of Mathematical Linguistics* (Vol. 101(1), pp. 7–28).

% of terms	EN-ES		EN-FR		EN-ET		EN-LT		EN-LV		EN-RU		EN-DE	
	B	I	B	I	B	I	B	I	B	I	B	I	B	I
Term correct	71.3	85.4	55.9	75.0	39.8	40.4	42.1	64.2	51.3	67.9	60.2	89.2	70.3	85.6
Wrong inflection	1.9	11.5	1.6	5.8	19.4	50.3	7.9	18.4	11.3	27.5	6.0	8.7	1.6	5.2
Not translated	8.6	0.6	19.2	14.3	9.9	1.2	0.6	0.0	4.6	0.0	16.3	0.9	7.8	0.3
Term split up or re-ordered	2.2	0.3	6.7	0.9	2.2	0.3	1.9	2.2	2.6	0.0	6.6	0.6	0.7	1.0
Wrong lexical choice	7.3	1.3	13.3	1.6	18.8	4.6	30.4	2.8	20.9	0.0	10.8	0.6	5.9	5.6
Not a term	6.4	0.6	1.7	1.7	0.6	0.6	10.8	10.8	1.7	1.7	0.0	0.0	0.7	1.0
Other	2.2	0.3	1.6	0.7	9.3	2.5	6.3	1.6	7.6	3.0	0.0	0.0	13.1	1.3
Rel. impr. of correct term translations (%)	19.6		34.1		1.6		52.6		32.3		48.0		21.9	
Rel. impr. of correct lexical choice (%)	32.2		40.5		53.1		65.2		52.4		47.7		26.4	
Rel. red. of errors (%)	48.9		43.3		1.0		38.3		34.0		72.7		51.6	

Table 5: Evaluation summary for term translation quality

Bertoldi, N., and Federico, M. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proceedings of the 4th Workshop on Statistical Machine Translation* (pp. 182189).

Bouamor, D., Semmar, N., and Zweigenbaum, P. 2012. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *Proceedings of LREC 2012* (pp. 674–679).

Carl, M., and Langlais, P. 2002. An Intelligent Terminology Database as a Pre-processor for Statistical Machine Translation. In *COLING-02 on COMPUTERM 2002: 2nd international workshop on computational terminology*. (Vol. 14, pp. 1–7).

Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Buck, C., and Germann, U. 2014. The MateCat Tool. In *Proceedings of COLING 2014* (pp. 129–132).

Hálek, O., Rosa, R., Tamchyna, A., and Bojar, O. 2011. Named Entities from Wikipedia for Machine Translation. In *Proceedings of ITAT 2011* (pp. 23–30).

Hildebrand, A. S., Eck, M., Vogel, S., and Waibel, A. 2005. Adaptation of the Translation Model for Statistical Machine Translation Based on Information Retrieval. In *Proceedings of EAMT* (pp. 133–142).

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., and Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Interactive Poster and Demonstration Sessions* (pp. 177–180).

Koehn, P., and Schroeder, J. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation* (pp. 224–227).

Nikoulina, V., Sandor, A., and Dymetman, M. 2012. Hybrid Adaptation of Named Entity Recognition for Statistical Machine Translation. In *Proceedings of ML4HMT-12* (pp. 116).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002* (pp. 311–318).

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., and Gornostay, T. 2012. Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. In *Proceedings of TKE 2012* (pp. 193–208).

Pinnis, M., and Skadiņš, R. 2012. MT Adaptation for Under-Resourced Domains What Works and What Not. In *Proceedings of Baltic HLT 2012* (pp. 177–180).

Randolph, J. J.. 2005. Free-Marginal Multirater Kappa (multirater K[free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. In *Joensuu Learning and Instruction Symposium*.

Skadiņš, R., Skadiņa, I., Pinnis, M., Vasiļjevs, A., and Hudík, T. 2014. Application of Machine Translation in Localization into Low-resourced Languages. In *Proceedings of EAMT 2014* (pp. 209–216).

Spärck Jones, K. 1972. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. In *Journal of Documentation* 28, 11–21.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlter, P. 2012. DGT-TM: A Freely Available Translation Memory in 22 Languages. In *Proceedings of LREC 2012* (pp. 454–459).

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., and Tiedemann, J. 2012. LetsMT!: a Cloud-Based Platform for Do-It-Yourself Machine Translation. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 43–48).