Domain Adaptation in Statistical Machine Translation Using Comparable Corpora: Case Study for English-Latvian IT Localisation

Mārcis Pinnis, Inguna Skadiņa, and Andrejs Vasiļjevs

Tilde
{marcis.pinnis,inguna.skadina,andrejs}@tilde.lv

Abstract. In the recent years, statistical machine translation (SMT) has received much attention from language technology researchers and it is more and more applied not only to widely used language pairs, but also to under-resourced languages. However, under-resourced languages and narrow domains face the problem of insufficient parallel data for building SMT systems of reasonable quality for practical applications. In this paper we show how broad domain SMT systems can be successfully tailored to narrow domains using data extracted from strongly comparable corpora. We describe our experiments on adaptation of a baseline English-Latvian SMT system trained on publicly available parallel data (mostly legal texts) to the information technology domain by adding data extracted from in-domain comparable corpora. In addition to comparative human evaluation the adapted SMT system was also evaluated in a real life localisation scenario. Application of comparable corpora provides significant improvements increasing human translation productivity by 13.6% while maintaining an acceptable quality of translation.

Keywords: comparable corpus, statistical machine translation, software localisation, under-resourced languages, Latvian, narrow domain.

1 Introduction

In the recent years, SMT has become the dominant paradigm not only for widely-used languages, but also for under-resourced languages. However, lack of sufficiently large parallel corpora limits the building of reasonably good quality machine translation (MT) solutions for these languages. Because of this reason there is a growing interest in research of comparable corpora as a source for extracting data useful for training MT systems.

In this paper we describe our research on using comparable corpora for adaptation of an SMT system for translation from English into the under-resourced language: Latvian. The Latvian language belongs to the Baltic language group of the Indo-European language family, with less than 2.5 million speakers worldwide. It is a morphology rich language with a rather free word order. Since there is a relatively small number of Latvian speakers, content in Latvian is also limited. Only few bi/multilingual parallel corpora contain Latvian, among them the largest are JRC-Acquis [21], DGT-TM [22], and Opus [24].

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 224–235, 2013.

These corpora have sufficient data only for building legal domain SMT systems with high BLEU scores when evaluated on in-domain texts [18]. However, these systems are not suitable for other domains, e.g., automotive or information technology (IT).

Although quality of MT systems has been criticized a lot, due to a growing pressure on efficiency and cost reduction, MT receives more and more interest from the localisation industry. Localization companies have to increase volume of translation and decrease costs of services in order to remain competitive in the market.

In this paper we address both these challenges. We show that, for language pairs and domains where there is not enough parallel data available (1) in-domain comparable corpora can be used to increase translation quality and (2) if comparable corpora are large enough and can be classified as strongly comparable (i.e., have many similar text fragments, sentence pairs or phrases overlapping between the different languages) then the trained SMT systems applied in the localisation process increase productivity of human translators.

In the next chapters we present our work on English-Latvian SMT system adaptation to the IT domain: building a comparable corpus, extracting semi-parallel sentences and terminological units from the comparable corpus, and adapting the SMT system to the IT domain with the help of the extracted data. We describe evaluation results demonstrating that data extracted from comparable corpora can significantly increase BLEU score over a baseline system. Results from the application of the adapted SMT system in a real life localisation task are presented showing that SMT usage increased the productivity of human translators by 13.6%.

2 Related Work

2.1 Comparable Corpora in Machine Translation

Applicability of comparable corpora for MT is a relatively new field of research. While methods on how to use parallel corpora in MT are well studied (e.g. [6]), methods and techniques for comparable corpora have not been thoroughly investigated.

The latest research has shown that adding extracted parallel lexical data from comparable corpora to the training data of a SMT system improves the system's performance in view of word coverage [5]. It has been also demonstrated that language pairs with little parallel data can benefit the most from exploitation of comparable corpora [8]. Munteanu and Marcu [9] achieved significant performance improvements from large comparable corpora of news feeds for English, Arabic and Chinese over a baseline MT system, trained on existing available parallel data.

However, most of such experiments are performed with widely used language pairs, such as French-English [1, 2], Arabic-English [2] or English-German [23], while for under-resourced languages (e.g., Latvian), possible exploitation of comparable corpora for machine translation tasks is less studied [17].

2.2 Machine Translation in Localisation

Different aspects of post-editing and machine translatability have been researched since the 90-ies (a comprehensive overview has been provided by O´Brien [11]). Recently several productivity tests have been performed in translation and localisation

industry settings at Microsoft [16], Adobe [4], Autodesk [15] and Tilde [19]. In all these tests authors report productivity increase. However, in many cases they also indicate on significant performance differences in the various translation tasks. Also increase of the error score for translated texts is reported.

As the localization industry experiences a growing pressure on efficiency and performance, some developers have already integrated MT in their computer-assisted translation (CAT) products, e.g. SDL Trados, ESTeam TRANSLATOR and Kilgrey memoQ.

3 Collecting and Processing Comparable Corpus

3.1 Comparable Corpus

For our experiment we used an English-Latvian comparable corpus containing texts from the IT domain: software manuals and Web crawled data (consisting of IT product information, IT news, reviews, blogs, user support texts including also software manuals, etc.). The corpus was acquired in an artificial fashion in order to simulate a strongly comparable narrow domain corpus (that is, a corpus containing overlapping content in a significant proportion).

To get more data for our experiments we used two different approaches in creation of comparable corpus. Thus the corpus consists of two parts. The first part contains documents acquired from different versions of software manuals of a productivity software suite split in chunks of less than 100 paragraphs per document and aligned at document level with *DictMetric* tool [20]. As a very large number of alignments was produced, we filtered document pairs so that for each source and target language document there were no more than the top three alignments (for both languages separately) included.

The second part consists of an artificially created strongly comparable corpus from parallel data that is enriched with Web crawled non-comparable and weakly comparable data. The parallel data was split in random chunks from 40 to 70 sentences per document and randomly polluted with sentences from the Web crawled data from 0 to 210 sentences. The Web corpus sentences were injected in random positions in English and Latvian documents separately, thus heavily polluting the documents with non-comparable data. The Web crawled data was collected using the *Focussed Monolingual Crawler* (FMC) from the ACCURAT Toolkit [12]. The Web corpus consists of 232,665 unique English and 96,573 unique Latvian sentences. The parallel data before pollution contained 1'257,142 sentence pairs.

The statistics of the English-Latvian comparable corpus are given in Table 1. Note that the second part of the corpus accounts for 22,498 document pairs.

Table 1. Comparable corpus statistics

English documents	Latvian documents	Number of aligned document pairs	Number of aligned document pairs after filtering
27,698	27,734	385,574	45,897

Although, this comparable corpus has been artificially created, the whole process chain of system adaptation described in the following sections is the same for any comparable corpus, e.g., it can be applied to corpora automatically acquired from the Web.

3.2 Extraction of Semi-parallel Sentence Pairs

The parallel sentence extractor *LEXACC* [23] was used to extract semi-parallel sentences from the comparable corpus. Before extraction, texts were pre-processed – split into sentences (one sentence per line) and tokenized (tokens separated by a space).

Because the two parts of our corpus differ in terms of comparable data distribution and the comparability level, different confidence score thresholds were applied for extraction. The threshold was selected by manual inspection of extracted sentences so that most (more than 90%) of the extracted sentence pairs would be strongly comparable or parallel.

Table 2 shows information about data extracted from both parts of the corpus using the selected thresholds.

Table 2. Extracted semi-parallel sentence pairs

Corpus part	Threshold	Unique sentence pairs
First part	0.6	9,720
Second part	0.35	561,994

3.3 Extraction of Bilingual Term Pairs

We applied the ACCURAT Toolkit to acquire in-domain bilingual term pairs from the comparable corpus following the process thoroughly described in [13], which then were used in the SMT adaptation process. At first, the comparable corpus was monolingually tagged with terms and then terms were bilingually mapped. Term pairs with the confidence score of mapping below the selected threshold were filtered out. In order to achieve a precision of about 90%, we selected the confidence score threshold of 0.7. The statistics of both the monolingually extracted terms and the mapped terms are given in Table 3.

Table 3. Term tagging and mapping statistics

Corpus part	Unique monolingual terms		Mapped term pairs	
	English	Latvian	Before filtering	After filtering
First part	127,416	271,427	847	689
Second part	415,401	2,566,891	3,501	3,393

The term pairs were further filtered so that for each Latvian term only those English terms having the highest mapping confidence scores would be preserved. We used Latvian term to filter term pairs, because Latvian is a morphologically richer language and multiple inflective forms of a word in most cases correspond to a single English word form (although this is a “rude” filter, it increases the precision of term mapping to well over 90%).

As can be seen in Table 3, only a small part of the monolingual terms were mapped. However, this amount of mapped terms was sufficient for SMT system adaptation as described in the following sections. It should also be noted that in our adaptation scenario translated single-word terms are more important than multi-word terms as the adaptation process of single-word terms partially covers also the multi-word pairs that have been missed by the mapping process.

4 Building SMT Systems

We used the LetsMT! platform [25] based on the Moses tools [7] to build three SMT systems: the baseline SMT system (trained on publicly available parallel corpora), the intermediate adapted SMT system (in addition data extracted from comparable corpus was used) and the final adapted SMT system (in-domain terms integrated). All SMT systems have been tuned with minimum error rate training (MERT) [3] using in-domain (IT domain) randomly selected tuning data containing 1,837 unique sentence pairs.

4.1 Baseline SMT System

For the English-Latvian baseline system, the DGT-TM parallel corpora of two releases (2007 and 2011) were used. The corpora were cleaned in order to remove corrupt sentence pairs and duplicates. As a result, for training of the baseline system a total of 1'828,317 unique parallel sentence pairs were used for translation model training and a total of 1'736,384 unique Latvian sentences were used for language model training.

4.2 Domain Adaptation through Integration of Data Extracted from Comparable Corpora

In order to adapt the SMT system for the IT domain, the extracted in-domain semi-parallel data (both sentence pairs and term pairs) were added to the parallel corpus used for baseline SMT system training. The whole parallel corpus was then cleaned and filtered with the same techniques as for the baseline system. The statistics of the filtered corpora used in SMT training of the adapted systems (intermediate and final) are shown in Table 4.

Table 4. Training data for adapted SMT systems

	Parallel corpus (unique pairs)	Monolingual corpus
DGT-TM (2007 and 2011) sentences	1'828,317	1'576,623
Sentences from comparable corpus	558,168	1'317,298
Terms form comparable corpus	3,594	3,565

Table 4 shows that there was some sentence pair overlap between the DGT-TM corpora and the comparable corpora content. This was expected as DGT-TM covers a broad domain and may contain documents related to the IT domain. For language modelling, however, the sentences that overlap in general domain and in-domain

monolingual corpora have been filtered out from the general domain monolingual corpus. Therefore, the DGT-TM monolingual corpus statistics between the baseline system and the adapted system do not match.

After filtering, a translation model was trained from all available parallel data and two separate language models were trained from the monolingual corpora:

- Latvian sentences from the DGT-TM corpora were used to build the general domain language model;
- The Latvian part of extracted semi-parallel sentences from in-domain comparable corpus were used to build the in-domain language model.

4.3 Domain Adaptation through Terminology Integration

To make in-domain translation candidates distinguishable from general domain translation candidates, the phrase table of the domain adapted SMT system was further transformed to a term-aware phrase table [14] by adding a sixth feature to the default five features used in Moses phrase tables. The following values were assigned to this sixth feature:

- "2" if a phrase in both languages contained a term pair from the list of extracted term pairs.
- "1" if a phrase in both languages did not contain any extracted term pair; if a phrase contained a term only in one language, but not in both, it received "1" as this case indicates of possible out-of-domain (wrong) translation candidates;

In order to find out whether a phrase contained a given term or not, every word in the phrase and the term itself was stemmed. Finally, the transformed phrase table was integrated back into the adapted SMT system.

5 Automatic and Comparative Evaluation

5.1 Automatic Evaluation

The evaluation of the baseline and both adapted systems was performed with four different automatic evaluation metrics: BLEU, NIST, TER and METEOR on 926 unique IT domain sentence pairs. Both, case sensitive and case insensitive, evaluations were performed. The results are given in Table 5.

Table 5. Automatic evaluation results

System	Case sensitive?	BLEU	NIST	TER	METEOR
Baseline	*No*	*11.41*	*4.0005*	*85.68*	*0.1711*
	Yes	10.97	3.8617	86.62	0.1203
Intermediate	*No*	*56.28*	*9.1805*	*43.23*	*0.3998*
adapted system	Yes	54.81	8.9349	45.04	0.3499
Final adapted	*No*	***56.66***	***9.1966***	***43.08***	***0.4012***
system	Yes	**55.20**	**8.9674**	**44.74**	**0.3514**

The automatic evaluation shows a significant performance increase of the improved systems over the baseline system in all evaluation metrics. Comparing two adapted systems, we can see that making the phrase table term-aware (*Final adapted system*) yields further improvement over intermediate results after just adding data extracted from comparable corpora (*Intermediate adapted system*). This is due to better terminology selection in the fully adapted system. As terms comprise only a certain part of texts, the improvement is limited.

5.2 Comparative Evaluation

For the system comparison we used the same test corpus as for automatic evaluation and compared the baseline system against the adapted system. Figure 1 summarizes the human evaluation results using the evaluation method described in [18]. From 697 evaluated sentences, in 490 cases (70.30±3.39%) output of the improved SMT system was chosen as a better translation, while in 207 cases (29.70±3.39%) users preferred the translation of the baseline system. This allows us to conclude that for IT domain texts the adapted SMT system provides better translations as the baseline system.

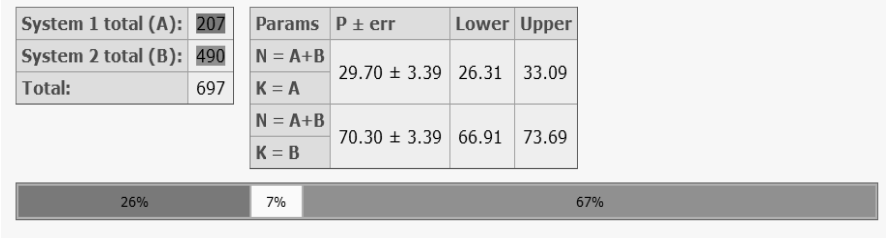

Fig. 1. System comparison by total points (System 1 – baseline, System 2 – adapted system)

Figure 2 illustrates the evaluation on sentence level: for 35 sentences we can reliably say that the adapted SMT system provides a better translation, while only for 3 sentences users preferred the translation of the baseline system. It must be noted that in this figure we present the results only for those sentences for which there was a statistically significant preference to the first or second system by the evaluators.

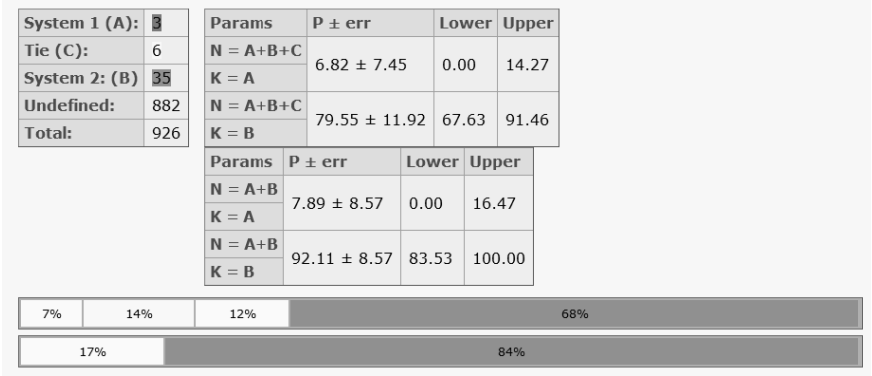

Fig. 2. System comparison by count of the best sentences (System 1 – baseline, System 2 – adapted system)

6 Evaluation in Localisation Task

The main goal of this evaluation task was to evaluate whether integration of the adapted SMT system in the localisation process allows increasing the output of translators in comparison to the efficiency of manual translation. We compared productivity (words translated per hour) in two real life localisation scenarios:

- Translation using translation memories (TM's) only.
- Translation using suggestions of TM's and the SMT system that is enriched with data from comparable corpus.

6.1 Evaluation Setup

For tests 30 documents from the IT domain were used. Each document was split into two parts. The length of each part of a document was 250 to 260 adjusted words on average, resulting in 2 sets of documents with about 7,700 words in each set.

Three translators with different levels of experience and average performance were involved in the evaluation cycle. Each of them translated 10 documents without SMT support and 10 documents with integrated SMT support. The SDL Trados translation tool was used in both cases.

The results were analysed by editors who had no information about techniques used to assist the translators. They analysed average values for translation performance (translated words per hour) and calculated an error score for translated texts. Individual productivity of each translator was measured and compared against his or her own productivity. An error score was calculated for every translation task by counting errors identified by an editor and applying a weighted multiplier based on the severity of the error type (1):

$$ErrorScore = \frac{1000}{n} \sum_i w_i e_i \quad (1)$$

where n is a number of words in translated text, e_i is a number of errors of type i, w_i is a coefficient (weight) indicating the severity of type i errors. Depending on the error score the translation gets a translation quality grade (*Superior*, *Good*, *Mediocre*, *Poor* or *Very poor*) assigned (Table 6).

Table 6. Quality grades based on error scores

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

6.2 Results

Usage of MT suggestions in addition to TM's increased the productivity of the translators on average from 503 to 572 words per hour (13.6% improvement). There were significant differences in the results of different translators from performance increase by 35.4% to decreased performance by 5.9% for one of the translators (see Table 7). Analysis of these differences requires further studies but most likely they are caused

by working patterns and the skills of individual translators. The average productivity for all the translators has been calculated using the formula (2).

$$Productivity\ (scenario) = \frac{\sum_{Text=1}^{N} Adjusted\ words(Text,scenario)}{\sum_{Text=1}^{N} Actual\ time(Text,scenario)} \quad (2)$$

Table 7. Results of productivity evaluation

Translator	Scenario	Actual productivity	Productivity increase or decrease	Standard deviation of productivity
Translator 1	TM	493.2	35.39%	110.7
	TM+MT	667.7		121.8
Translator 2	TM	380.7	13.02%	34.2
	TM+MT	430.3		38.9
Translator 3	TM	756.9	-5.89%	113.8
	TM+MT	712.3		172.0
Average	**TM**	**503.2**	**13.63%**	**186.8**
	TM+MT	**571.9**		**184.0**

According to the standard deviation of productivity in both scenarios (without MT support 186.8 and with MT support 184.0) there were no significant performance differences in the overall evaluation (see Table 8). However, each translator separately showed higher differences in translation performance when using the MT translation scenario.

The overall error score (shown in Table 8) increased for one out of three translators. Although the total increase in the error score for all translators combined was from 24.9 to 26.0 points, it still remained at the quality evaluation grade "Good".

Table 8. Localisation task error score results

Translator	Scenario	Accuracy	Language quality	Style	Termino-logy	Total error score
Translator 1	TM	6.8	8.0	6.8	1.6	**23.3**
	TM+MT	9.9	14.4	7.8	4.1	**36.3**
Translator 2	TM	8.2	10.1	11.7	0.0	**30.0**
	TM+MT	3.8	11.7	7.6	1.5	**24.6**
Translator 3	TM	4.6	9.5	7.3	0.0	**21.4**
	TM+MT	3.0	8.3	6.0	0.8	**18.1**
Average	**TM**	**6.5**	**9.3**	**8.6**	**0.5**	**24.9**
	TM+MT	**5.4**	**11.4**	**7.1**	**2.1**	**26.0**

7 Conclusion

The results of our experiment demonstrate that it is feasible to adapt SMT systems for a particular domain with the help of comparable data and integrate such SMT systems for highly inflected under-resourced languages into the localisation process.

The use of the English->Latvian domain adapted SMT suggestions (trained on comparable data) in addition to the translation memories lead to the increase of translation performance by 13.6% while maintaining an acceptable ("*Good*") quality of the translation. However, our experiments also showed a relatively high difference in translator performance changes (from -5.89% to +35.39%), which suggests that for more justified results the experiment should be carried out with more participants. It would also be useful to further analyse correlation between the regular productivity of translator and the impact on productivity by adding MT support.

Error rate analysis shows that overall usage of MT suggestions decreased the quality of translation in two error categories (language quality and terminology). At the same time this degradation is not critical and the result is still acceptable for production purposes.

To our knowledge, this is the first evaluation of usability of SMT systems enriched with comparable data for translation into a less-resourced highly inflected language. This is also one of the first evaluation of SMT for an under-resourced highly inflected language in the localisation environment.

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. Many thanks to our colleagues Juris Celmiņš, Elita Kalniņa and Artūrs Pudulis for participation in the localisation experiments and Ieva Dātava for her support and comments.

References

1. Abdul-Rauf, S., Schwenk, H.: On the use of comparable corpora to improve SMT performance. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, pp. 16–23 (2009)
2. Abdul-Rauf, S., Schwenk, H.: Parallel sentence generation from comparable corpora for improved SMT. Machine Translation 25(4), 341–375 (2011)
3. Bertoldi, N., Haddow, B., Fouet, J.B.: Improved Minimum Error Rate Training in Moses. The Prague Bulletin of Mathematical Linguistics 91, 7–16 (2009)
4. Flournoy, R., Duran, C.: Machine translation and document localization at Adobe: From pilot to production. In: MT Summit XII: Proceedings of the Twelfth Machine Translation Summit, Ottawa, Canada (2009)
5. Hewavitharana, S., Vogel, S.: Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. In: Proceedings of the Workshop on Comparable Corpora, LREC 2008, pp. 7–10 (2008)
6. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2010)

7. Koehn, P., Federico, M., Cowan, B., Zens, R., Duer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, pp. 177–180 (2007)
8. Lu, B., Jiang, T., Chow, K., Tsou, B.K.: Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. In: Proceedings of the 3rd Workshop on Building and using Comparable Corpora: from Parallel to Non-Parallel Corpora, Valletta, Malta, pp. 42–48 (2010)
9. Munteanu, D., Marcu, D.: Extracting parallel sub-sentential fragments from nonparallel corpora. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Morristown, NJ, USA, pp. 81–88 (2006)
10. Munteanu, D., Marcu, D.: Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Computational Linguistics 31(4), 477–504 (2006)
11. O´Brien, S.: Methodologies for measuring the correlations between post-editing effort and machine translatability. Machine Translation 19(1), 37–58 (2005)
12. Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., Babych, B.: ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In: Proceedings of System Demonstrations Track of ACL 2012, Jeju Island, Republic of Korea (2012)
13. Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., Gornostay, T.: Term Extraction, Tagging and Mapping Tools for Under-Resourced Languages. In: Proceedings of the 10th Conference on Terminology and Knowledge Engineering, Madrid, Spain (2012)
14. Pinnis, M., Skadiņš, R.: MT Adaptation for Under-Resourced Domains – What Works and What Not. In: Proceedings of the Fifth International Conference Baltic HLT 2012, pp. 176–184. IOS Press, Tartu (2012)
15. Plitt, M., Masselot, F.: A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. The Prague Bulletin of Mathematical Linguistics 93, 7–16 (2010)
16. Schmidtke, D.: Microsoft office localization: use of language and translation technology (2008), http://www.tm-europe.org/files/resources/TM-Europe2008-Dag-Schmidtke-Microsoft.pdf
17. Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlič, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M.L., Pinnis, M.: Collecting and Using Comparable Corpora for Statistical Machine Translation. In: Proceedings of LREC 2012, Istanbul, Turkey, May 21-27, pp. 438–445 (2012)
18. Skadiņš, R., Goba, K., Šics, V.: Improving SMT for Baltic Languages with Factored Models. In: Proceedings of the Fourth International Conference Baltic HLT 2010, Riga, Latvia, October 7-8, pp. 125–132 (2010)
19. Skadiņš, R., Puriņš, M., Skadiņa, I., Vasiļjevs, A.: Evaluation of SMT in localization to under-resourced inflected language. In: Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011, Leuven, Belgium, May 30-31, pp. 35–40 (2011)
20. Su, F., Babych, B.: Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-) Parallel Translation Equivalents. In: Proceedings of the EACL 2012 Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), Avignon, France, April 23-27, pp. 10–19 (2012)

21. Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, P.: DGT-TM: A freely Available Translation Memory in 22 Languages. In: Proceedings of LREC 2012, Istanbul, Turkey, pp. 454–459 (2012)
22. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of LREC 2006, Genoa, Italy, pp. 2142–2147 (2006)
23. Ştefănescu, D., Ion, R., Hunsicker, S.: Hybrid Parallel Sentence Mining from Comparable Corpora. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), Trento, Italy, May 28-30, pp. 137–144 (2012)
24. Tiedemann, J.: News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: Recent Advances in Natural Language Processing, vol. V, pp. 237–248 (2009)
25. Vasiļjevs, A., Skadiņš, R., Tiedemann, J.: LetsMT!: a cloud-based platform for do-it-yourself machine translation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Jeju, Republic of Korea, July 10. System Demonstrations, pp. 43–48 (2012)