Chapter 6
Training, Enhancing, Evaluating and Using MT Systems with Comparable Data

Bogdan Babych, Yu Chen, Andreas Eisele, Sabine Hunsicker, Mārcis Pinnis, Inguna Skadiņa, Raivis Skadiņš, Gregor Thurmair, Andrejs Vasiļjevs, Mateja Verlic, and Xiaojun Zhang

Abstract This chapter describes how semi-parallel and parallel data extracted from comparable corpora can be used in enhancing machine translation (MT) systems: what are the methods used for this task in statistical and rule-based machine translation systems; what kinds of showcases exist that illustrate the usage of such enhanced MT systems. The impact of data extracted from comparable corpora on MT quality is evaluated for 17 language pairs, and detailed studies involving human evaluation are carried out for 11 language pairs. At first, baseline statistical machine translation (SMT) systems were built using traditional SMT techniques. Then they were improved by the integration of additional data extracted from the comparable corpora. Comparative evaluation was performed to measure improvements. Comparable corpora were also used to enrich the linguistic knowledge of rule-based machine translation (RBMT) systems by applying terminology extraction technology. Finally, SMT systems were adjusted for a narrow domain and included domain-specific knowledge such as terminology, named entities (NEs), domain-specific language models (LMs), etc.

Chapter editor: Inguna Skadiņa

B. Babych
University of Leeds, Leeds, UK
e-mail: b.babych@leeds.ac.uk

Y. Chen · A. Eisele · S. Hunsicker · X. Zhang
Deutsche Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Saarbrücken, Germany

M. Pinnis · I. Skadiņa (✉) · R. Skadiņš · A. Vasiļjevs
Tilde, Riga, Latvia
e-mail: inguna.skadina@tilde.lv

G. Thurmair
Linguatec, München, Germany

M. Verlic
Zemanta, Ljubljana, Slovenia

I. Skadiņa et al. (eds.), *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*, Theory and Applications of Natural Language Processing,
https://doi.org/10.1007/978-3-319-99004-0_6

6.1 Introduction

Building a statistical machine translation (SMT) system requires a large amount of parallel data for model training. Reasonably good results can be achieved when the domain of the training corpus is close to the test data.

There are only a few parallel corpora publicly available for the lesser spoken languages of Europe. Several large-scale highly multi-lingual parallel language resources, such as the JRC-Acquis corpus (Steinberger et al. 2006), the DGT-TM (Steinberger et al. 2012) and DCEP corpus (Najeh et al. 2014), are made available by the European Commission's Joint Research Centre (JRC) and other European Union organisations (Steinberger et al. 2014). Different corpora are presented in the OPUS collection (Tiedemann 2009, 2012). SETimes (Tyers and Alperen 2010) is a parallel corpus from a multi-lingual news website into English and eight South-East European Languages (Albanian, Bulgarian, Croatian, Greek, Macedonian, Romanian, Serbian and Turkish).

For many under-resourced languages, multi-lingual comparable resources are widely available. Data extracted from comparable resources can be useful for machine translation. While methods on how to use parallel corpora in MT are well studied, methods and techniques for comparable corpora have not been thoroughly investigated.

The research in the field of the application of comparable corpora to the task of SMT has shown that adding extracted aligned parallel lexical data (additional phrase tables and their combination) from comparable corpora to the training data of an SMT system improves the system's performance in view of untranslated word coverage (Hewavitharana and Vogel 2008; Xu et al. 2006; Zhang 2011). It has also been demonstrated that language pairs with little parallel data can benefit the most from exploitation of comparable corpora (Lu et al. 2010).

Xu et al. (2006) exploit comparable data to extract parallel corpus. The proposed approach breaks documents into segments using pre-defined anchor words and then align these segments. In order to avoid errors in alignments, they present an advanced approach to extract the parallel sentences recursively by partitioning a bilingual document into two pairs. For Chinese–English data, this method produced translation results comparable to those of a state-of-the-art sentence aligner. A combination of the two approaches lead to a better translation performance.

Munteanu and Marcu (2006) 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. The authors stated that the impact of comparable corpora on SMT performance is 'comparable to that of human translated data of similar size and domain'.

Irvine and Callison-Burch (2013) used comparable corpora to improve accuracy and coverage of phrase-based MT built on small amounts of parallel data. They showed that adding translations of low-frequency words from comparable corpora improves performance beyond what is achieved by inducing translations for out-of-

vocabulary words alone and that data from comparable corpora improves BLEU score (Papineni et al. 2002).

Most of the experiments are performed with widely used language pairs, such as French–English (Abdul-Rauf and Schwenk 2009, 2011), Arabic–English (Abdul-Rauf and Schwenk 2011) or English–German (Ştefănescu et al. 2012), while possible exploitation of comparable corpora for machine translation tasks is less studied for under-resourced languages (e.g. Skadiņa et al. 2012).

In this chapter, we analyse the impact of data extracted from comparable corpora on the machine translation task (both data-driven and rule-based) for under-resourced languages and narrow domains. Section 6.2 describes experiments to improve SMT systems trained on available parallel data by integration of additional data from comparable corpora for application in the general domain translation task. Section 6.3 proposes a methodology for how to assess changes in translation quality for systems enhanced with data extracted from comparable corpora and describes human evaluation results for eleven language pairs. Section 6.4 focusses on MT adaptation for a particular domain with the help of domain data extracted from comparable corpora. The last three sections deal with use cases. Section 6.5 analyses German–English MT adaptation to the automotive domain for both (rule-based and SMT) approaches. Section 6.6 analyses the role of machine translation in Web authoring, while Sect. 6.7 discusses the application of MT systems, enriched with data from comparable corpora, in computer-aided translation.

6.2 Enriching General Domain SMT Systems with Data from Comparable Corpora

In this section, we describe experiments to improve SMT systems trained on available parallel data (we call them baseline systems) by integration of additional data from comparable corpora for application in the general domain translation task.

6.2.1 Data Used for Experiments

The following publicly accessible parallel corpora were used to set up baseline SMT systems for the experiments: JRC: JRC-Acquis, DGT: DGT-TM (Steinberger et al. 2012), SETimes,[1] Europarl, and News Commentary.[2] Table 6.1 shows the size of the training data that was used to train the baseline systems.

[1] http://www.setimes.com

[2] The News Commentary corpus is from the training data released for the shared tasks of the last few workshops for statistical machine translation (SMT).

Table 6.1 Size of corpora for baseline systems

Language Pair	Corpora	Size (lines)
English–Latvian	DGT, JRC	2,305,674
English–Lithuanian	DGT, JRC	2,339,905
English–Estonian	DGT, JRC	2,239,791
English–Slovenian	DGT, JRC	2,190,704
German–Romanian	DGT, JRC	615,336
Latvian–Lithuanian	DGT, JRC	974,161
Lithuanian–Romanian	DGT, JRC	940,461
English–Greek	SETimes	169,337
English–Croatian	SETimes	157,950
English–Romanian	SETimes	171,573
Greek–Romanian	SETimes	175,019
German–English	Europarl, Newscommentary	1,639,893

We conducted three groups of directions in our experiments. The first group uses JRC and DGT for training and the second group uses SETimes. Although the data combining JRC and DGT is fairly large in size, the domain of the data is rather limited to legislation/law. The systems based on such a data set perform poorly on general translation tasks of other open domains, in spite of the high translation quality for in-domain tests reported in previous literature. Therefore, we still consider these language pairs under-resourced. The second group is the opposite. This group of baseline systems is based on the SETimes corpus, which covers a relatively broad range of topics and is much smaller in size than JRC or DGT. The third group includes only German–English as a control group. We used both Europarl and News Commentary for this group. This dataset has a presumably open domain and large size. This setup allows us to have more contrastive studies on the effect of using comparable corpora, as the set up for German–English has been used for state-of-the-art systems.

As for language model (LM) training, we use the target portion of the corresponding parallel data.

To enrich the baseline SMT systems, we use data extracted from comparable corpora collected by tools described in Chap. 3. We distinguish between the data extracted from news corpora (News) and Wikipedia articles corpora (Wiki).

The ACCURAT toolkit (Pinnis et al. 2012a) was used to extract semi-parallel sentences from the aligned comparable corpora. Table 6.2 gives the statistics about the extracted data. The amount of data varies a lot between language pairs and also between the two comparable corpora.

We used the News corpus to adapt the language models. The amount of data is reported in Table 6.3.

We tune all models on the same development set (Table 6.4) to get comparable results. The tuning is performed using minimal error rate training (MERT; Och 2003).

Table 6.2 Statistics of the extracted semi-parallel data from comparable corpora

	Number of lines	
Language Pair	News	Wiki
English–Latvian	112,398	116,240
English–Lithuanian	33,219	179,578
English–Estonian	19,048	128,939
English–Slovenian	67,508	5418
German–Romanian	10,227	–
Latvian–Lithuanian	7163	29,370
Lithuanian–Romanian	9470	–
English–Greek	6641	45,646
English–Croatian	36,663	31,048
English–Romanian	23,820	45,771
Greek–Romanian	1783	–
German–English	13,782	–

Table 6.3 Statistics of monolingual comparable corpora

Language	Size (lines)
Croatian	180,908
German	1,485,774
Greek	1,267,731
English	2,235,282
Estonian	711,147
Latvian	789,178
Lithuanian	1,061,713
Romanian	1,815,170
Slovenian	470,782

Table 6.4 Statistics about development data

Language Pair	Name of development set	Length (in lines)
English–Latvian	Tilde	1000
English–Lithuanian	Tilde	1000
English–Estonian	Tilde	1000
English–Greek	SETimes	600
English–Croatian	SETimes	600
Croatian–English	SETimes	600
English–Romanian	SETimes	600
Romanian–English	SETimes	600
English–Slovenian	mtserver	1000
Slovenian–English	mtserver	1000
German–English	WMT-dev 2008	2051
German–Romanian	RACAI	3000
Romanian–German	RACAI	3000
Greek–Romanian	SETimes	600
Romanian–Greek	SETimes	600
Lithuanian–Romanian	DGT-dev	3000
Latvian–Lithuanian	Tilde	1000

Additionally, we make use of the target language tuning texts to interpolate the language models as described in the next subsection.

6.2.2 Methodology

When improving SMT systems, we need to look at the two models used in translation: the translation model (TM) and the language model (LM). The comparable data can be used to adapt both models.

6.2.2.1 Mixture Translation Model

Including additional parallel corpora as training data to an SMT system usually yields an improvement to a certain extent. However, the additional texts could also introduce errors that do not exist in the original model. This case is especially more likely to happen when the parallel texts are not translations of each other: for example when we have misaligned sentences in the comparable corpora. On the other hand, due to various reasons, the added data might not be dominant enough among the other sources of training corpora to help the SMT system to recover from the errors in the baseline system. Therefore, in addition to a single translation model built from both the parallel corpora and the comparable data as a whole, we experimented with mixture models that distinguish texts from different sources.

The mixture models, introduced by Xu et al. (2007), start from individual models that are generated separately using the sets of texts from different sources. The most straightforward way is to divide the data into two subsets: the original parallel corpora versus the aligned texts that were extracted from the comparable corpus. Such a partition may be very close to the baseline model when the sizes of the two subsets differ too much, as it would lead to a mixture model that relies on the larger subset. Thus, in order to emphasise and better control the contribution of parallel and comparable data to the final translation, we choose to further divide the original parallel data into separate corpora, from each of which we generate a different translation model. This approach also allows us to understand the influence of each individual corpus (parallel or comparable) in the SMT system, and it is especially important when the parallel corpora used in the baseline systems are from very different domains.

As a state-of-the-art word alignment algorithm such as GIZA++ tends to perform poorly for a limited amount of data, we generate the word alignments for the mixture model by training over the combination of all the training data, that is the parallel data alongside the extracted sentence pairs from the comparable corpus in order to find sufficient alignment points that are useful for constructing a translation model. Then, after the second step, the word alignments are split into segments corresponding to the individual corpus.

We construct the individual translation models from the word alignments for each corpus. The models are then sorted by the size of the corresponding training corpora, given the fact that the probabilistic estimation over a larger set of data is usually more reliable.

The other models are appended to the largest model in this sorted order such that only phrase pairs that were never seen previously are included. Lastly, we add new features (in the form of additional columns) to the phrase table of the final translation model to indicate each phrase pair's origin. Each new column corresponds to one model, including the original model. If a phrase table entry appears in a model, its feature value in the corresponding column is 2.718; otherwise, it is 1.

Table 6.5 shows a few sample entries from the phrase table of a mixture model created in our experiments for English–Latvian translation. The first five columns are the probabilistic scores estimated in the standard phrase-based SMT training, including the inverse phrase translation probability $\varphi(f|e)$, the inverse lexical weighting $lex(f|e)$, the direct phrase translation probability $\varphi(e|f)$, the direct lexical weighting $lex(e|f)$ and the phrase penalty which is always $e^1 = 2.718$. Following the scheme of defining the phrase penalty, we added three additional columns to the phrase table, corresponding to the three individual models which have been sorted by size. In this example, the first column refers to the JRC model, the second column refers to the DGT model and the last column is for the extracted comparable corpus. The values in these three columns are either 2.718 or 1, indicating whether the phrase pairs exist in the individual models. For example the last three columns for the phrase pair '*economic approaches*'-'*ekonomiskas metodes*' are 1, 2.718, and 1. This means that this pair is originally from the DGT model and does not appear in the other two.

In the mixture model, segments repeated by many sources are considered more probable for translation. On the other hand, unique pieces from some sources may lead us to valuable information, such as terminologies from a particular domain in the comparable corpus. The former case corresponds to phrase pairs with very high probabilities, whereas the latter is still included in the model.

Table 6.5 Sample entries from the phrase table of a mixture model for English–Latvian

Source phrase (e)	Target phrase (f)	Probabilistic scores	Origin markers
Economic, political	Ekonomiskās, politiskās	0.079 0.266 0.011 0.011 2.718	2.718 2.718 2.718
Economic, social	Ekonomiskajā, sociālajā	0.119 0.048 0.001 0.001 2.718	2.718 2.718 1
Economic downturn	Ekonomikas lejupslīdi	0.120 0.134 0.017 0.016 2.718	1 2.718 2.718
Economic subjects	Ekonomiskajos priekšmetos	0.406 0.555 0.051 0.001 2.718	1 2.718 1
Economic approaches	Ekonomiskas metodes	0.241 0.004 0.241 0.001 2.718	1 2.718 1

6.2.2.2 Interpolating Language Models

To make the best use of the fact that our language models have been trained on different texts, we want to combine them into one and adapt the n-gram probabilities accordingly. Although, for example, our baseline JRC and DGT language models are out of domain, we do not want to completely lose the information they contain. On the other hand, these models are big enough that they can overpower the influence of the new language model that has been trained on much smaller amounts of data. Here we need to adjust the n-gram probabilities so that they mirror what we would expect from our target domain.

Combination is done by optimising the perplexity of the interpolated language model on an in-domain development text in the target language. We then receive a lambda for each language model we used; we can adjust the probabilities for each n-gram. In this way, we combine the probabilities from the different language models into one (Schwenk and Koehn 2008).

The interpolated language model will then be used for the new SMT system.

6.2.3 Experiments with Data Extracted from Comparable Corpora

In total, we worked on seventeen language pairs: English–Latvian, English–Lithuanian, English–Estonian, English–Greek, English–Croatian, Croatian–English, English–Romanian, Romanian–English, English–Slovenian, Slovenian–English, German–English, German–Romanian, Romanian–German, Greek–Romanian, Romanian–Greek, Lithuanian–Romanian, Latvian–Lithuanian. Our main concern is to translate from English, but we also investigate a few language pairs that do not involve English and for which there is very little data available.

We trained state-of-the-art phrase-based models using 7-gram phrase-tables and 5-gram interpolated language models. For the training, we used the data described in Table 6.1, where the parallel data was used for the translation model and the target language text was used to generate the language model. In the case of the language pairs using DGT and JRC, as well as German–English, we interpolated the language models built on the two baseline corpora using the target side of our development set. This is the same set that we later optimised the SMT translation parameters on using Minimal Error Rate Training (MERT) and is listed in Table 6.4.

Then, for each language pair, we trained systems using the additional data described in Table 6.2. We use the same general settings for training the enriched models as we did for training the baseline models. We trained separate models for the data extracted from the News and the Wiki data to examine the influence of the different sorts of data.

For the interpolated model, we use the target side of both the baseline parallel data and the collected comparable corpus. The translation model is trained on the

extracted parallel data and the baseline corpora. We apply this approach to both the News and the Wiki extracted data. For the language model, we use the comparable News corpus for both News and Wiki experiments.

For the mixture model, we trained a phrase table on each individual corpus and then combined them into a single mixture translation model. For the language model, we used the interpolated language models.

All systems were tested on the same test set, which consists of 511 sentences from general domain text (Skadiņš et al. 2010). Table 6.6 lists the results for all experiments on interpolated language models and mixture models. Figures in bold indicate models that outperform the baseline. The best model for each language pair is denoted with an asterisk.

We see that not every approach works equally well for each language direction. The largest improvement in BLEU score can be noted for those language pairs that only used the SETimes corpus with less than 200,000 lines per language pair as the baseline corpus. The improvements are smaller for the language pairs using DGT/JRC. For some of the language pairs, we did not observe any improvement by adding the data, and thus, we further investigated English–Lithuanian pair. We describe these experiments in the next subsection.

Table 6.6 Evaluation results (BLEU scores) for all experiments

Language pair	Baseline	Interpolated LM		Mixture models
		News	Wiki	
English–Latvian	12.74	**13.20** (+.46)	**13.07**(+.33)	**13.25*** (+.51)
English–Lithuanian	12.66	12.21(−.45)	12.33 (−.33)	11.94 (−.71)
English–Estonian	10.44	**11.23*** (+.79)	**10.46** (+.02)	**10.88** (+.44)
English–Greek	19.06	**21.40** (+2.34)	**23.67* (+4.61)**	**20.61** (+1.55)
English–Croatian	10.91	10.36 (−.55)	**11.25** (+.34)	**11.45*** (+.54)
Croatian–English	20.78	20.31 (−.47)	**21.17** (+.39)	**21.91***(+1.13)
English–Romanian	17.89	**20.11*** (+2.22)	**20.00** (+2.11)	**19.08** (+1.19)
Romanian–English	21.54	**26.16** (+4.62)	**30.35*** (+8.81)	**25.27** (+3.73)
English–Slovenian	18.20	**18.68*** (+.48)	**18.66** (+.46)	17.70 (−.50)
Slovenian–English	26.28	**27.40** (+1.12)	**27.46*** (+1.18)	**27.31** (+1.03)
German–English	27.90	**28.62*** (+.72)	–	27.88 (−.02)
German–Romanian	9.66	**10.14*** (+.48)	–	8.37 (−1.29)
Romanian–German	10.22	9.56 (−.66)	–	9.97 (−.25)
Greek–Romanian	15.81	**17.25*** (+1.44)	–	**17.15** (+1.34)
Romanian–Greek	12.13	**13.59*** (+1.46)	–	**13.37** (+1.24)
Lithuanian–Romanian	9.91	9.24 (−.67)	–	4.67 (−5.24)
Latvian–Lithuanian	12.12	**12.69*** (+.57)	8.70 (−3.42)	**12.41** (+.29)

6.2.4 Staggered Experiments

The LEXACC tool, which is described in Chap. 5, assigns a score to each sentence pair extracted from comparable corpora, denoting how likely these two sentences are parallel. As such, the LEXACC score should allow us to predict how usable a particular chunk of data is, that is, will the use of this data increase translation quality.

To test this influence of the LEXACC score, we split up the extracted data. We want to check the effect of the score both in intervals and in a cumulative fashion. The hypothesis for the former is that data with a higher LEXACC score should be more helpful than data with a lower score. In the cumulative experiments, we choose different thresholds. As the score goes down, the less parallel the data will become, and more errors will be introduced into the translation model. But as the distribution of the data follows Zipf's law, we have very few items with a very high score, but, the lower the score, the more sentences LEXACC extracts. However, we also need to take into account how much data we have: for higher thresholds, LEXACC will only be able to extract small amounts of data. Here we are interested in the threshold that allows the maximal increase in translation quality for the amount of data used. This threshold may vary for different corpora which is an effect we also want to examine.

As we couldn't observe an improvement in translation quality in the experiments using the full data for English–Lithuanian, we treat this language in these experiments. Additionally, we examine English–Latvian and English–Romanian. We saw improvements in these two languages, but we are interested in seeing how much each part of the data contributes. We chose these languages, because they work with different baseline corpora. This allows us to see the effects of adding a small amount of data to a large out-of-domain corpus (DGT/JRC in the case of English–Latvian) and the effects of adding similar amounts of data to a small in-domain corpus (SETimes for English–Romanian).

6.2.4.1 English–Latvian

For English–Latvian, we examined both the interpolated language models and the mixture models. The problem with using mixture models is that the probabilities associated with the entries in the phrase table become less trustworthy on such a small set of data. Tables 6.7 and 6.8 give the amount of data (in sentence pairs) in the different intervals.

We did not investigate data with an LEXACC score of less than 0.1 (the default threshold of LEXACC is 0.1). We see that we have very little data with a score higher than 0.9, but we get more data for lower scores.

We used each chunk of the data to retrain the SMT model and evaluated it the same as the baseline and full enriched models. Tables 6.9 and 6.10 give the BLEU scores for those experiments. The baseline SMT system reached a BLEU score of 12.66. Experiments that perform worse than the baseline are marked in italics; the best experiment in each approach and corpus is marked in boldface.

Table 6.7 Statistics about interval experiments for English–Latvian

Interval	News	Wiki
>0.9	169	208
0.9–0.8	3226	1730
0.8–0.7	13,264	5791
0.7–0.6	12,735	6868
0.6–0.5	9009	7085
0.5–0.4	6914	8556
0.4–0.3	8720	13,902
0.3–0.2	15,325	26,669
0.2–0.1	43,036	45,431

Table 6.8 Statistics about cumulative experiments for English–Latvian

Cumulative	News	Wiki
>0.9	169	208
>0.8	3395	1938
>0.7	16,659	7729
>0.6	29,394	14,597
>0.5	38,403	21,682
>0.4	45,317	30,238
>0.3	54,037	44,140
>0.2	69,362	70,809
>0.1	112,398	116,240

Table 6.9 BLEU scores for interval experiments for English–Latvian

	Interpolated LM		Mixture models	
Interval	News	Wiki	News	Wiki
>0.9	13.48	13.73	12.97	13.48
0.9–0.8	13.60	13.57	**13.29**	13.36
0.8–0.7	13.15	13.57	12.71	13.29
0.7–0.6	**13.67**	**13.83**	12.76	13.23
0.6–0.5	13.49	13.50	12.84	12.91
0.5–0.4	13.54	13.57	12.78	**13.72**
0.4–0.3	13.31	13.39	12.80	13.61
0.3–0.2	12.77	13.40	12.99	13.44
0.2–0.1	*12.15*	*12.63*	12.84	12.86

Figure 6.1 illustrates the effect of the LEXACC score on the BLEU score. The data in the interval of [0.1,0.2] scores the worst results and doesn't even reach the BLEU score of the baseline (plotted for comparison purposes). As the LEXACC score increases, we can also see an increase in BLEU score. Using the interpolated language models, this development is rather steady. When we compare News to the Wiki-extracted data, the interpolated language models show similar trends.

According to the BLEU scores, the translation results using the mixture models seem less correlated to the LEXACC score, mostly due to the fact that the mixture models are very sensitive to the size of the data that is used to construct the additional phrase tables. Higher LEXACC thresholds indicate better quality of extracted sentence pairs. Meanwhile, these high scores also result in less extracted data. In

Table 6.10 BLEU scores for cumulative experiments for English–Latvian

	Interpolated LM		Mixture models	
Cumulative	News	Wiki	News	Wiki
>0.9	13.48	**13.73**	12.97	13.48
>0.8	13.50	13.34	13.77	12.90
>0.7	13.66	*12.56*	13.19	**13.49**
>0.6	**13.86**	13.55	**13.78**	12.97
>0.5	13.73	13.10	13.00	13.11
>0.4	13.68	13.30	13.41	12.90
>0.3	13.58	13.22	13.26	12.96
>0.2	13.74	13.46	13.75	13.15
>0.1	13.20	13.07	13.25	–

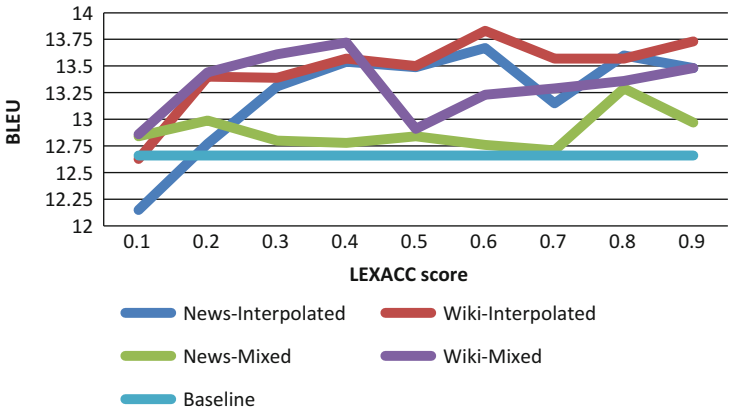

Fig. 6.1 BLEU scores for interval experiments for English-Latvian

general, the translation model constructed over a small amount of data tends to contain less useful phrase pair entries while having high probability estimation values. When combining a small model with high scores and a much larger model with much lower scores, one cannot avoid penalising the phrase pairs from the small model in order to use entries that exist in the other models, which are actually the majority of the combined model. Thus, in general, the tuning procedure seems to assign higher weights to the feature that represents the larger model. As a result, the additional data might not have as much influence on the final translation as we hope. It also explains why, in the experiment for Wiki data, the BLEU score drops significantly at the LEXACC interval [0.4,0.5], for which there are nearly 40% less sentence pairs than for [0.3,0.5]. The BLEU score increases again for higher LEXACC scores, as the size difference is smaller for the other cases. In practice, the probability estimation in the sub-models should all be normalised, but this would make it more difficult to compare results for different extracted data. Therefore, we chose to retain the probability scores in the sub-models.

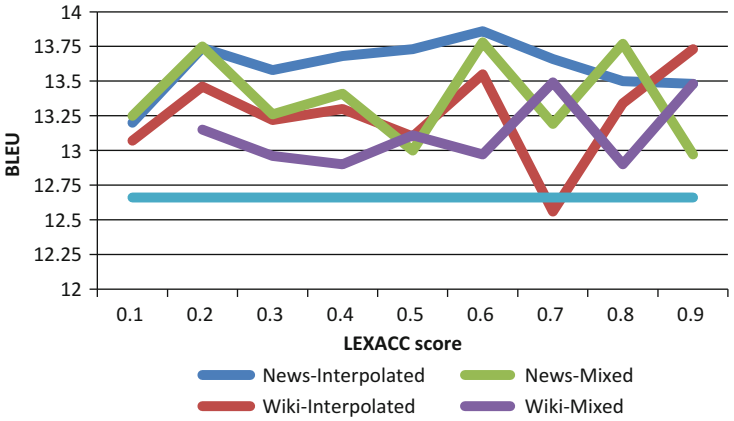

Fig. 6.2 BLEU scores for cumulative experiments for English–Latvian

The results for the cumulative experiments are not quite as clear. The effect of the LEXACC score on BLEU is plotted in Fig. 6.2. Here we see a lot of fluctuation. Although the best BLEU scores are comparable for three of the four experiment runs, they occur in different intervals. Especially interesting is the behaviour of the data with an LEXACC score of 0.7 and above. In News, this chunk leads to an improvement using the interpolated LMs, but the BLEU score drops by almost 0.6 for the mixture models which is a significant deterioration. The Wiki data behaves similarly, except that here the BLEU score of the interpolated LM drops even below the baseline performance. However, this data is the best performing for the mixture models

Figure 6.2 illustrates this point. We see a lot of ups and downs, although the data using a threshold of 0.6 seems to work reliably well for both models and both corpora.

6.2.4.2 English–Romanian

The training data for English–Romanian was very small, so our hypothesis was that this language direction was very sensitive to the quality of the newly added data. Whereas the DGT/JRC corpora are big enough to smooth out mistakes in the translation probabilities, the SETimes corpus is small enough that even the relatively small amount of extracted data can counteract the probabilities extracted from the original data: the English–Latvian baseline corpus consists of 2,305,674 lines, with 112,398/116,240 lines extracted from each comparable corpus, adding about 5% of the data to the baseline corpus. For English–Romanian, we only had 171,573 lines in the baseline, so the data from News (238,320 lines) and the Wiki corpus (45,771 lines) amount to 14% and 27%, respectively. Thus, the influence of the new data will be much higher than for the previous experiments.

Table 6.11 Statistics about interval experiments for English–Romanian

Interval	News	Wiki
>0.9	246	5807
0.9–0.8	2468	13,174
0.8–0.7	2221	6530
0.7–0.6	1511	3993
0.6–0.5	2021	3653
0.5–0.4	2636	3974
0.4–0.3	4024	3826
0.3–0.2	8693	4814

Table 6.12 Statistics about cumulative experiments for English–Romanian

Cumulative	News	Wiki
>0.9	246	5807
>0.8	2714	18,981
>0.7	4935	25,511
>0.6	6446	29,504
>0.5	8467	33,157
>0.4	11,103	37,131
>0.3	15,127	40,957
>0.2	23,820	45,771

For this language pair, we examined only the interpolated language models, as the results on the mixture models were too unsteady. Tables 6.11 and 6.12 give the amount of data in the different intervals.

The distribution of this data is especially interesting. In English–Latvian, the distribution followed Zipf's law, that is there was very little data for the high scores, but the lower the score, the more data was extracted. For English–Romanian, however, this only holds for News. The Wiki corpus behaves differently: here we have unusually many sentence pairs with a high score. This cannot simply be explained by the fact that Wiki articles are inherently more strongly comparable than news text, as then this would also have to hold for other language pairs. Manual inspection of the data suggests that many articles in the Romanian Wikipedia have been originally translated from the English Wikipedia. We consider this an anomaly.

The procedure of these experiments is the same as for the previous English–Latvian experiments. For each chunk of the data, we retrain the SMT models and compare it against the baseline, which was evaluated with a BLEU score of 17.89. Table 6.13 shows the results for the interval experiments, the best results are marked in boldface.

All systems outperform the baseline, but the overall tendency for improvement of BLEU is not as clear-cut as it was for the previous experiment (Fig. 6.3). Instead, we see that the improvement in BLEU varies a lot over of the intervals. For the Wiki corpus, which adds 25% to the original data, our assumption that higher LEXACC scores predict a higher increase in BLEU still holds, but, for the News data, we find that using the maximum amount of available data results in the highest gain. Here we must take into account the amount of data in each interval: although Wiki can offer

Table 6.13 BLEU scores for interval experiments for English–Romanian

Interval	Interpolated LM	
	News	Wiki
>0.9	19.45	19.08
0.9–0.8	18.74	19.63
0.8–0.7	19.28	**19.64**
0.7–0.6	19.81	18.79
0.6–0.5	20.03	19.13
0.5–0.4	20.04	19.29
0.4–0.3	**20.22**	19.30
0.3–0.2	19.92	18.30

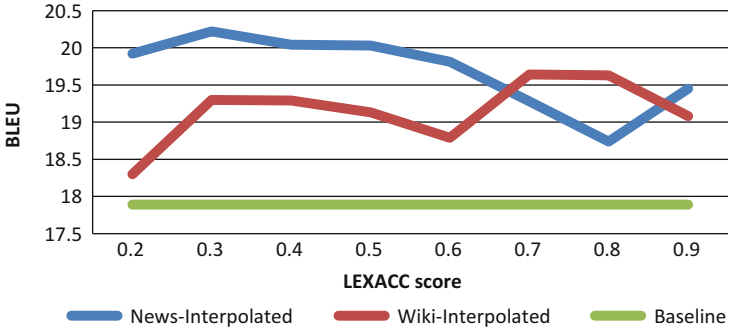

Fig. 6.3 BLEU scores for interval experiments for English–Romanian

Table 6.14 BLEU scores for cumulative experiments for English–Romanian

Cumulative	Interpolated LM	
	News	Wiki
>0.9	19.45	19.08
>0.8	19.04	19.59
>0.7	18.54	19.75
>0.6	18.71	20.03
>0.5	19.01	19.98
>0.4	19.85	**20.27**
>0.3	19.44	20.40
>0.2	**20.11**	20.00

us 13,000 additional lines in the interval of [0.9,0.8], there are only 2500 sentences in the same interval in the News corpus.

Table 6.14 shows the results for the cumulative experiments, the best results are marked in boldface. As for the interval experiments, all models improve over the baseline.

In Fig. 6.4, we see less variation than for English–Latvian, with rather obvious thresholds for the corpora. As for the interval experiments, we get the best results by using all of the available additional data for the News corpus, whereas the threshold for Wiki lies at 0.3. This is consistent with the best LEXACC performance, where we reach the best F1 score at a threshold of 0.36. Although these thresholds are close, we

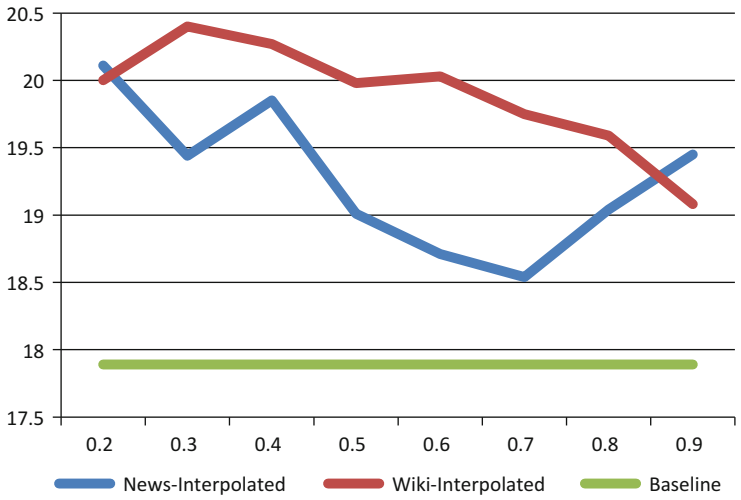

Fig. 6.4 BLEU scores for cumulative experiments for English–Romanian

see quite a difference between the different corpora: the News corpus improves by 0.7 BLEU points when using all the data, whereas the performance of the Wiki corpus drops by 0.3 BLEU points when using the same threshold. The BLEU scores for threshold 0.3 differ by almost one full BLEU score, a very significant difference. This can be explained by taking into account the amount of data (see Table 6.12); for this interval, we have almost three times as many sentences for Wiki than for News.

6.2.4.3 English–Lithuanian

As shown in Table 6.6, using the full data for English–Lithuanian did not result in an improvement of BLEU score. As we have seen a lot of variation in the BLEU scores for the individual chunks of the data, we decided to give English–Lithuanian the same treatment.

The size of the original baseline corpus consisting of DGT/JRC was 2,339,905 lines. We could add 33,219 lines to this from the News corpus (+1.42%) and 179,578 lines from Wiki (+7.67%). Splitting up the data into the individual chunks results in the amount of data shown in Tables 6.15 and 6.16.

The data again follows the distribution we would expect. The difference in size between the News and Wiki corpus is significant—in each section, we have about six times as much data for Wiki than for the News corpus.

The baseline produced a BLEU score of 12.66. Tables 6.17 and 6.18 present the BLEU scores for the respective interval and cumulative experiments, the best results are marked in boldface.

None of the interval experiments perform better than the baseline, but we can see that the Wiki data performs much better than the News data. In Fig. 6.5, we observe the general tendency that higher scoring intervals result in better BLEU scores, but

Table 6.15 Statistics about interval experiments for English–Lithuanian

Interval	News	Wiki
>0.9	28	1089
0.9–0.8	352	4265
0.8–0.7	1006	6450
0.7–0.6	1061	6307
0.6–0.5	1317	7656
0.5–0.4	1692	10,393
0.4–0.3	2495	17,628
0.3–0.2	5536	35,574
0.2–0.1	19,732	90,196

Table 6.16 Statistics about cumulative experiments for English–Lithuanian

Cumulative	News	Wiki
>0.9	28	1089
>0.8	380	5354
>0.7	1386	11,804
>0.6	2447	18,111
>0.5	3764	25,767
>0.4	5456	36,160
>0.3	7951	53,788
>0.2	13,487	89,562
>0.1	33,219	179,758

Table 6.17 BLEU scores for interval experiments for English–Lithuanian

	Interpolated LM	
Interval	News	Wiki
>0.9	**12.48**	**12.64**
0.9–0.8	12.00	12.49
0.8–0.7	12.47	12.40
0.7–0.6	12.47	12.53
0.6–0.5	12.33	12.37
0.5–0.4	12.46	12.00
0.4–0.3	12.01	12.26
0.3–0.2	12.04	12.34
0.2–0.1	12.13	11.87

the amount of data does not seem sufficient to push the enriched system over the baseline.

Using the interval, especially the small amounts available for the News corpus, did not yield an improvement in the system.

Most of the cumulative experiments also perform worse than the baseline (Fig. 6.6). It is interesting to note that the best-performing system, which also improves over the baseline, uses the same threshold we have already identified as optimal for English–Latvian, namely 0.6. This can be interpreted such that Lithuanian generally behaves similar to Latvian.

Table 6.18 BLEU scores for cumulative experiments for English–Lithuanian

	Interpolated LM	
Cumulative	News	Wiki
>0.9	12.48	**12.64**
>0.8	12.35	12.56
>0.7	12.35	12.34
>0.6	**12.94**	12.43
>0.5	11.90	12.41
>0.4	12.11	12.32
>0.3	12.45	12.25
>0.2	12.37	11.93
>0.1	11.21	12.33

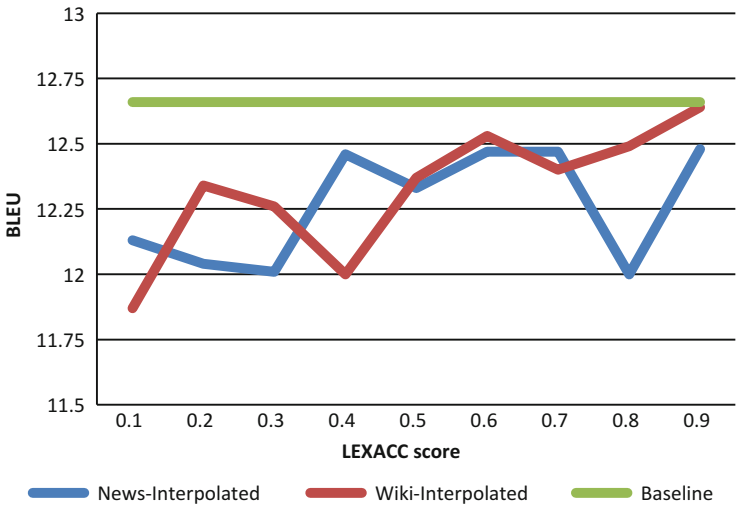

Fig. 6.5 BLEU scores for interval experiments for English–Lithuanian

It is worthwhile to note that the upper intervals get close to the performance of the baseline which leads us to believe that the amount of data extracted was simply too small to have a large enough impact on the baseline corpus.

6.3 Human Evaluation of MT Output

The human evaluation experiment is designed to measure the difference between the performance of the baseline MT systems built using only parallel data and those that were enhanced with sentences and phrases extracted from comparable corpora (CC). We developed a special evaluation scenario which takes into account the properties of the evaluated data.

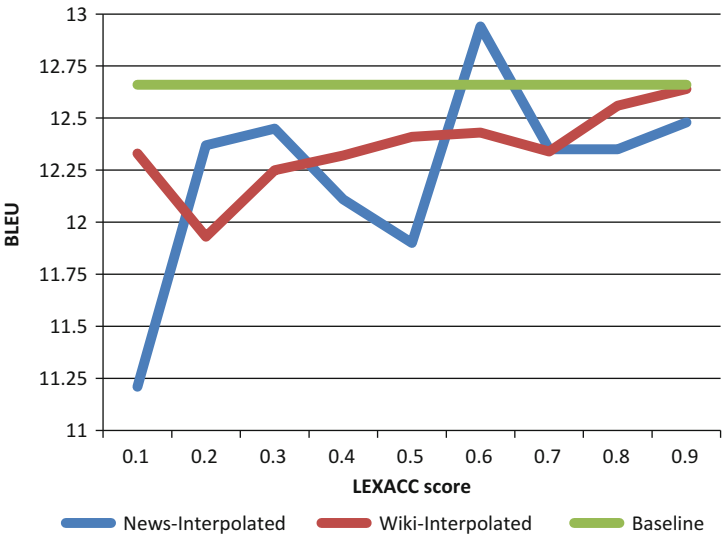

Fig. 6.6 BLEU scores for cumulative experiments for English–Lithuanian

Traditional measures of translation quality involve metrics such as adequacy (fidelity, i.e. the amount of information preserved in MT output compared with the gold-standard human translation), fluency (the degree of naturalness or well-formedness of a sentence according to the requirements of the target language, irrespective of the original sentence) or informativeness (responses to a multiple-choice questionnaire; White et al. 1994). However, we noted that none of these standard measures can adequately quantify the changes in translation quality for systems enhanced with resources based on comparable corpora.

6.3.1 Evaluation Methodology and the Interface

For our evaluation experiment, we developed a novel evaluation methodology, which captures differences between the baseline and the modified MT systems more directly and systematically. Specifically, we were interested in the following aspects of MT evaluation.

Firstly, we need to capture a general user intuition about the translation quality of evaluated sentences taken in context. The division between adequacy and fluency makes sense for non-translator users; however, our target audience—translation studies students or professional translators—are able to assess the relevant importance of adequacy and fluency for their specific post-editing or summarisation tasks. In this respect, it makes sense to collapse both evaluation measures onto a single scale, for which we obtain professional user ratings. In our scenario, translators were

asked to evaluate the *overall translation quality* of the sentences presented to them in the order that they normally appear in a text.

Secondly, we are also interested in a comparative aspect of evaluation, specifically—the differences between the baseline sentences compared with corresponding aligned CC-enhanced sentences. Traditional comparative-based metrics have two major shortcomings from this perspective: they do not place compared sentences onto any systematic scale, and they do not compare specific linguistic: for example lexical differences within the sentences. In our case, we need to tie the differences to an interpretable scale and focus the attention of evaluators on specific changes in otherwise similar sentences. In our scenario, not all sentences are different in the baseline and the enhanced output, and, if there are differences, they are usually minimal; it can be just one or two words or different morphological forms of words. We also cannot use here standard adequacy or fluency measures independently on the baseline and CC-enhanced MT output; this would miss such small differences, since granularity of the standard 5-point scale could be insufficient for capturing the changes.

Therefore, in our evaluation scenario, we combined the question about the general translation quality with the comparative evaluation task: lexical differences between the baseline and the enhanced versions were highlighted, and users were asked to rate the appropriateness of lexical choices for each of the highlighted words. Sentences without any differences were removed (which sometimes disrupted the intra-sentential context, but the number of such omission was small compared to the overall text size), and the order of presentation was randomised. The origin of the text was anonymised; users did not know whether the sentence was coming from the baseline or the CC-enhanced MT system.

Highlighting lexical differences is intended to focus the attention of evaluators on specific linguistic issues, and the numerical scale combined with the comparative framework allows us to adequately quantify the quality level, as well as relative and absolute improvement in translation quality.

The evaluation interface presented to users had the following form (Fig. 6.7).

6.3.2 Experiment Set-Up

System output was generated for the baseline and CC-enhanced MT systems for the following translation directions and domains: News domain: German–English (de-en); Romanian–English (ro-en); Slovenian–English (sl-en); Croatian–English (hr-en); Romanian–German (ro-de); Latvian–Lithuanian (lv-lt); English–Latvian (en-lv); English–Croatian (en-hr); English–Greek (en-el); German–Romanian (de-ro); Greek–Romanian (el-ro); Automotive domain: German–English (de-en) and English–Latvian (en-lv). An evaluation set of 511 sentences (circa 11,000 words) was used for all translation directions.

Evaluation packs for human evaluation were constructed using the following procedure: sentences different in the baseline versus CC-enhanced output were identified. Words different in the baseline versus CC-enhanced output were

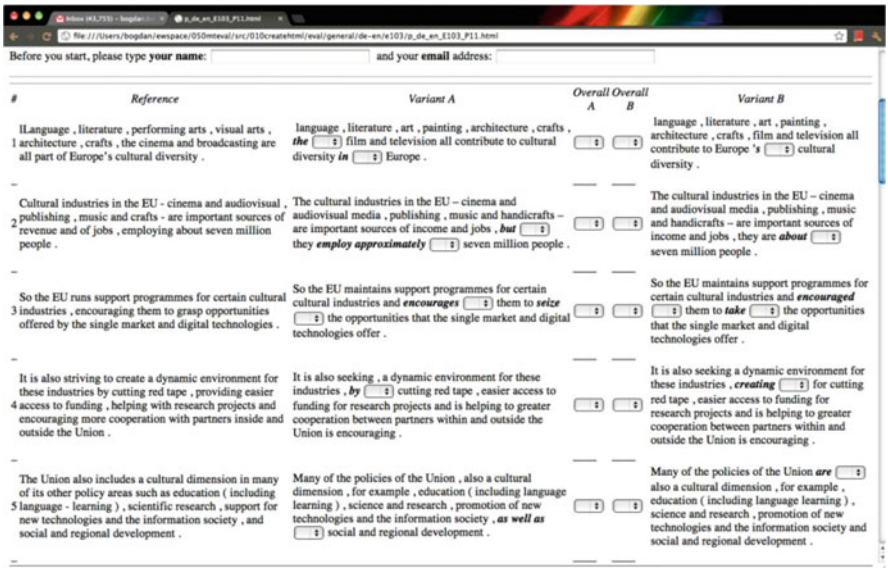

Fig. 6.7 Evaluation interface for professional translators

automatically highlighted; if several consecutive words were highlighted, all of them were evaluated together as a phrase. The order of presentation of the CC-enhanced versus baseline systems was randomised. Evaluation packs were presented to evaluators within a Web interface that automatically calculated submitted evaluation results (using CGI script).

The set of 120 sentences (those were the first 120 non-similar sentences out of the complete set of 511 sentences used for calculating BLEU scores) were typically used for the human evaluation experiment, with at least three independent judgments collected for each sentence and also—for each highlighted word or phrase that was different in the baseline versus CC-enhanced translation.

For each target language, we recruited at least three evaluators, most of them had backgrounds in translation (either professional translators, translation students or linguists), and obtained at least three independent scores for each of the compared sentences and lexical differences. Evaluators were asked to evaluate translation quality of the compared sentences and the quality of translation choices of the highlighted words or phrases.

For judging *overall quality,* evaluators were asked to rate each pair of sentences on a scale of 1 to 5: (1 = Translation is not at all good ... 5 = Translation is very good). For judging *lexical translation* choices, the judges were asked to use the same scale for rating translation quality of highlighted words and phrases: (1 = Very bad translation choice... 5 = Very good translation choice).

6.3.3 Human Evaluation Results

Evaluation results are presented in two groups: overall evaluation results (Table 6.19) and evaluation results for lexical differences (Table 6.20). Bold values denote the best results.

It can be seen from the table that improvement in translation quality is not observed for all translation directions. Even if the CC-enhanced system contains all the parallel data that is also present in the baseline MT, the addition of new comparable resources may cause degradation. The reason for this is that the data does not contain true translation equivalents which gives rise to spurious and wrong translations. These cases are less visible to automatic metrics like BLEU but are easily identified by human translators. Therefore, the important point is not only to be able to add more data to the system, but to control the quality of the data which is coming from comparable corpora.

Overall, the results show that the baseline translation quality is very low (1 or 2 on the 5-point translation quality scale on average). The quality of lexical translation choices (where translators were asked to focus on specific words or phrases that are different in the baseline and the enhanced output) is higher. Also, here CC-enhanced systems achieve greater improvement. This shows that the proposed evaluation methodology of focussing on lexical differences is more appropriate to the task of measuring improvements with CC-based data.

On average, across all translation directions, there is improvement in all four areas. The improvement was the smallest for overall translation quality in the

Table 6.19 Evaluation results for *overall translation quality*

Language pair	Baseline (average)	CC-enhanced	Human scores for improvement (%)
News			
de–en	2.269	2.1	–7.45
ro–en	**1.826**	**2.721**	**49.01**
sl–en	**1.869**	**2.025**	**8.35**
hr–en	2.175	2.199	1.10
ro–de	**1.692**	**1.846**	**9.10**
lv–lt	2.157	2.095	–2.87
en–lv	2.04	1.993	–2.30
en–hr	2.107	1.864	–11.53
en–el	2.212	2.362	6.78
de–ro	1.942	1.914	–1.44
el–ro	2.156	2.271	5.33
		Average	4.92
Automotive			
de–en	2.201	2.893	31.44
en–lv	2.177	2.5	14.84
		Average	23.14

Table 6.20 Evaluation results for *lexical choices*: baseline vs CC-enhanced MT

Language pair	Baseline	CC-enhanced	Human scores for improvement (%)
News			
de–en	2.773	2.774	0.04
ro–en	**1.819**	**3.377**	**85.65**
sl–en	2.642	2.867	8.52
hr–en	**2.66**	**2.905**	**9.21**
ro–de	2.351	2.376	1.06
lv–lt	2.614	2.587	−1.03
en–lv	2.564	2.618	2.11
en–hr	2.507	2.118	−15.52
en–el	3.026	3.271	8.10
de–ro	2.399	2.365	−1.42
el–ro	**2.757**	**3.458**	**25.43**
		Average	11.10
Automotive			
de–en	2.628	3.835	45.93
en–lv	2.604	2.956	13.52
		Average	29.72

News domain: 4.92% over the baseline, then lexical improvement in the News domain was 11.1% on average.

In the automotive domain, there is a much higher and consistent improvement for both evaluated systems and in both aspects: overall and lexical quality, in comparison to the broad domain. The improvement for automotive domains was 23.14% and 29.72% for overall and lexical translation quality, respectively.

For the broader News domain, improvement or deterioration depends on the translation direction. Translation into English is always improved. All cases of degradation are for translation into more morphologically complex languages, such as Croatian. The mechanism for this fact is not known and requires further investigation. The results point out that the biggest benefit of CC-enhanced data is achieved for narrow domains and for MT into morphologically simpler languages like English.

6.4 MT Adaptation for Under-Resourced Domains

This section focusses on a very practical aspect of statistical machine translation (SMT)—how a general out-of-domain SMT system can be tailored to a particular domain using data extracted from an in-domain comparable corpus. Particularly, we are dealing with domain-specific terminology and named entities (NEs). We extract terms and named entities from initial parallel training data. These terms and named entities are used to collect a comparable corpus from the Web. Then, we extract

parallel terms from the collected comparable corpus, and finally, we integrate them in the SMT system. The changes in the quality of the adapted SMT system are evaluated in respect to a general out-of-domain baseline system. This section is based on the publication by Pinnis and Skadiņš (2012).

6.4.1 Initial Extraction and Alignment of Terms and Named Entities

The first step in our SMT system adaptation technique is acquisition of in-domain term pairs. Bilingual terminology will allow making the SMT system term-aware and will allow finding better translation candidates for narrow-domain translation tasks. To acquire the term pairs, we use bilingual comparable corpora from the Web.

In order to find important domain specific documents on the Web, we use the small amount of available parallel data sentences (up to two or three thousand parallel sentences) and extract seed terms and named entities for a focussed narrow domain Web crawl. Terms and named entities are monolingually tagged in the parallel in-domain data. For terms, we use the *Tilde's Wrapper System for CollTerm* (*TWSC*) (Pinnis et al. 2012b) and for named entities—*TildeNER* (Pinnis 2012) for Latvian and *OpenNLP*[3] for English. In parallel, a *Moses* phrase table is created from the in-domain parallel data.

Then, the monolingually tagged terms and NEs (in our experiment, 542 unique English and 786 unique Latvian units in total) are bilingually aligned using the *Moses* phrase table. At first, we try to find all symmetric term and named entity phrases in the phrase table that have been monolingually tagged in both languages. We allow only full phrase table entry and term or named entity alignments; that is, a phrase is considered valid only if all tokens from the phrase are identical to tokens of the corresponding term or named entity. In order to also allow inflective form alignments, all tokens of all terms, named entities and phrases are stemmed prior to alignment. This allows finding more translation candidates in cases when some inflective forms have not been tagged as terms, but others have.

Then, we also align terms and named entities that have been tagged by only one of the monolingual taggers. If a phrase is aligned in the phrase table with multiple phrases from the other language, we select the translation candidate that has the highest averaged (source-to-target and target-to-source) translation probability within the phrase table. This step allows finding terms and NEs, which have been missed by one of the monolingual taggers, thus increasing the amount of extracted term and named entity phrases. The alignment method on the in-domain parallel data produced 783 bilingually aligned term and NE phrases.

[3] Apache OpenNLP (available at: http://opennlp.apache.org/).

6.4.2 Comparable Corpora Collection

The second step in our SMT system adaptation technique requires collection of bilingual in-domain comparable corpora from the Web. We use the bilingual terms and NEs that were extracted from the parallel in-domain data as seed terms for focussed monolingual crawling of two monolingual narrow domain Web corpora with the *Focussed Monolingual Crawler* (FMC), which is described in Chap. 3. By using bilingually aligned seed terms, we ensure that the crawled corpora will be comparable and in the same domain for both English and Latvian languages. As the aligned seed terms may also contain out-of-domain or cross-domain term and NE phrases, we apply a ranking method based on reference corpus statistics; more precisely, we use the inverse document frequency (IDF) (Spärck Jones 1972) scores of words from general (broad) domain corpora (e.g. the whole Wikipedia and current news corpora) to weigh the specificity of a phrase. We rank each bilingual phrase using the following equation:

$$R(p_{\text{src}}, p_{\text{trg}}) = \min\left(\sum_{i=1}^{|p_{\text{src}}|} \text{IDF}_{\text{src}}(p_{\text{src}}(i)), \sum_{j=1}^{|p_{\text{trg}}|} \text{IDF}_{\text{trg}}(p_{\text{trg}}(j))\right), \tag{6.1}$$

where p_{src} and p_{trg} denote phrases in the source and target languages and IDF_{src} and IDF_{trg} denote the respective language IDF score functions that return an IDF score for a given token. The ranking method was selected through a heuristic analysis so that specific in-domain term and named entity phrases would be ranked higher than broad-domain or cross-domain phrases. This technique also allows filtering out phrase pairs where a phrase may have a more general meaning in one language but a specific meaning in the other language. After applying a threshold on the ranks, 614 phrase pairs were kept in the seed term list for corpora collection.

In addition to the seed terms, FMC requires seed URLs. In total, 55 English and 14 Latvian in-domain seed URLs were manually collected.

When the seed terms and seed URLs were acquired, a 48-hour focussed monolingual Web crawl was initiated for both languages. The collected English and Latvian corpora were filtered for duplicates, broken into sentences, and tokenised. The statistics of the collected corpora are given in Table 6.21.

Both monolingual corpora were aligned in the document level using the *DictMetric* (Su and Babych 2012) tool described in Chap. 2, which scores document pair comparability and aligns document pairs that exceed a specified comparability score threshold. Executing *DictMetric* on narrow domain comparable corpora may cause over-generation of document pairs; that is, every document from one language can be paired with many documents from the other language. Therefore, we filtered the document alignments so that each Latvian document would be paired with the top three comparable English documents and vice versa, thus creating 81,373 document pairs. The comparable corpus statistics after document level alignment are given in Table 6.22.

Table 6.21 Monolingual automotive domain corpora statistics

Language	Unique documents	Sentences	Tokens	Unique sentences	Tokens in unique sentences
English	34,540	8,743,701	58,526,502	1,481,331	20,134,075
Latvian	6155	1,664,403	15,776,967	271,327	4,290,213

Table 6.22 English-Latvian automotive comparable corpus statistics

Language	Unique documents	Unique sentences	Tokens in unique sentences
English	24,124	1,114,609	15,660,911
Latvian	5461	247,846	3,939,921

6.4.3 *Extraction of Term Pairs from Comparable Corpus*

Once the bilingual comparable corpus is collected, the third step is to extract translated term pairs. Both parts (the Latvian and the English documents), similarly as in the first step, are monolingually tagged with *TWSC*. In this step, we only tag terms as the precision of named entity mapping without a phrase table is well below 90% and would create unnecessary noise in the extracted data for SMT adaptation. Then, by using the document alignment information of the comparable corpus, we map terms bilingually using the *TerminologyAligner* (*TEA*) (Pinnis et al. 2012b) tool with a translation confidence score threshold of 0.7 (with a precision of 90% and higher). In total, 369 in-domain term pairs were extracted from the bilingual comparable corpus.

6.4.4 *Baseline System Training*

We start with the creation of an English–Latvian baseline system using the following data:

- A relatively large out-of-domain parallel corpus. We used the publicly available DGT-TM (Steinberger et al. 2012) English-Latvian parallel corpus (release of 2007). The corpus consists of 804,501 unique parallel sentence pairs and 791,144 unique Latvian sentences. The Latvian part is used for language modelling.
- A small amount of in-domain parallel sentences (up to two or three thousand parallel sentences). In our experiments, we have selected the automotive domain (more precisely, service manuals) as the target domain. The in-domain data are split in two sets—tuning and evaluation. The tuning set and the evaluation set consist of 1745 and 872 unique sentence pairs from the automotive domain. All systems were tuned with minimum error rate training (MERT, Bertoldi et al. 2009) using the in-domain tuning set and evaluated on the evaluation set.

Table 6.23 Baseline system evaluation results

Case sensitive	BLEU	NIST	TER	METEOR
No	10.97	3.9355	89.75	0.1724
Yes	10.31	3.7953	90.40	0.1301

For MT system training, we use the *LetsMT!* (Vasiļjevs et al. 2012) Web-based platform for SMT system creation. The *LetsMT!* platform is built upon the state-of-the-art *Moses* (Koehn et al. 2007) SMT *Experiment Management System* (*EMS*).

Evaluation results for the baseline system using different automatic evaluation methods (BLEU (Papineni et al. 2002), NIST (Doddington 2002), TER (Snover et al. 2006), and METEOR (Banerjee and Lavie 2005)) are given in Table 6.23.

6.4.5 SMT System Adaptation

Following domain adaptation methods suggested in earlier research (Koehn and Schroeder 2007; Lewis et al. 2010; Xu et al. 2007),we start the SMT adaptation task by adding an in-domain language model built using the Latvian monolingual comparable corpus that was collected in the second step. We built the SMT system (named *Int_LM*) using two language models (a general and an in-domain model). Both language models have different weights determined with system tuning (MERT). The in-domain monolingual language model increases SMT quality to 11.3 BLEU points (a relative increase of only 3.0% over the baseline system). We also trained an SMT system (named *In-domain_LM_only*) using only the in-domain language model. The experiment achieved 11.16 BLEU points, which is an increase over the baseline system but also a decrease over the *Int_LM* system. This was expected, as MERT has tuned the in-domain language model to be more important, while the in-domain language model may not contain some general language phrases that are in the broad domain corpus (thus, also interpolation of the two models achieves a higher score).

We continue our experiments by adding the translated term pairs (in total 610) that were extracted from the in-domain tuning set to the parallel data corpus and the corresponding Latvian translations to the in-domain monolingual corpus, from which the SMT system is trained. This simple addition of in-domain term translations to the SMT system (named *Int_LM+T_Terms*) increased the quality to 12.93 BLEU points (a relative increase of 17.8% over the baseline system). After also adding term pairs extracted from the comparable corpus collected from the Web (in total 369 new pairs), the quality of the system (named *Int_LM+T&CC_Terms*) increased to 13.5 BLEU points (a relative increase of 23.1% over the baseline system).

Considering also term banks as possible translated term resources, we extracted 6767 unique in-domain automotive term pairs from EuroTermBank (Rirdance and Vasiljevs 2006).[4] Then, we trained an SMT system (named *Int_LM+ETB_Terms*)

[4]EuroTermBank (http://www.eurotermbank.com/).

with the same parameters as the *Int_LM+T_Terms* system. The system achieved 11.26 BLEU points, which is a decrease in comparison with the *Int_LM* system and much worse than *Int_LM+T&CC_Terms* (the best thus far performing system). The reason for the decrease is fairly simple—term banks, in many cases, provide multiple translation candidates for a single term. This causes ambiguities in the translation model and can result in the selection of the wrong translation hypothesis. To solve this issue (at least partially), the term pairs from the term bank would have to be semantically disambiguated in respect to the required domain so that only the correct in-domain pairs would be used in the SMT system training.

Recent results in MT system adaptation (Ştefănescu et al. 2012) suggest that pseudo-parallel sentence pairs extracted from in-domain comparable corpora and used for SMT system training can significantly improve SMT system quality. Using the same pseudo-parallel sentence extraction tool LEXACC, we extracted 6718 and 678 unique sentence pairs with two parallelism confidence score thresholds of 0.51 and 0.35. These sentence pairs were then added to the available parallel data and the in-domain monolingual corpus. The results after training the SMT systems (named *Int_LM+LEXACC_0.35* and *Int_LM+LEXACC_0.51*) show a decrease in BLEU points (10.75 and 11.08 respectively) in comparison with the *Int_LM* system. After manual analysis of the MT output for *Int_LM+LEXACC_0.35* in comparison with the baseline system, it was evident that the translation quality has decreased because of non-parallel sentence alignments in the LEXACC extracted sentence pairs that cause in-domain term phrase pairs to receive lower weights (translation probability scores) in the translation model. Although in-domain terms in the pseudo-parallel sentences are in many cases paired with correct translations, they are often also paired with incorrect translations, thus creating noise for the translation model. This is not to say that the pseudo-parallel sentences in general do not help to improve SMT quality but rather that, for very narrow and under-resourced domains, where it is difficult to find strongly comparable in-domain corpora in the Web, the results can lower translation quality because of incorrect term translation hypothesis.

So far in our experiments, only the in-domain language model helps to distinguish in-domain translation hypotheses from broad (general) domain hypotheses. Therefore, in the next step, we transformed the *Moses* phrase table of the translation model to an in-domain term-aware phrase table. We do this by adding a sixth feature to the default 5 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 list. If a phrase contains a term on one side (in one language) but not on the other, it receives the value ‘1’, as such situations indicate about possible out-of-domain (wrong) translation candidates.
- ‘2’ if a phrase contains a term pair from the term list on both sides (in both languages).

In order to find out whether a phrase in the phrase table contains a given term or not, phrases and terms are stemmed prior to comparison. This allows finding inflected forms of term phrases even if those are not given in the bilingual term

list. The sixth feature identifies phrases containing in-domain term translations and allows filtering out out-of-domain (wrong) translation hypotheses in the translation process.

With the described methodology, we transformed phrase tables of the systems *Int_LM+T_Terms* (using the 610 tuning data term pairs) and *Int_LM+T&CC_Terms* (additionally using the 369 term pairs from the comparable corpora) to term-aware phrase tables. After tuning with MERT, two new systems were created. The *Int_LM+T_Terms+6th* system achieves 13.19 BLEU points, and the *Int_LM+T&CC_Terms+6th* system achieves 13.61 BLEU points (a relative increase of 24.1% over the baseline system and the highest measured increase in this experiment). Although the increase in translation quality over the systems without the 6th feature is relatively small, the translations show better translation hypothesis selection for in-domain terminology.

Complete results of the previously described automotive domain systems are shown in Table 6.24 ('*CS*' stands for '*Case-Sensitive*' evaluation).

To show that improvements in SMT quality are also consistent when using larger corpora, we trained a new English–Latvian baseline system (*Big_Baseline*) using 5,363,043 parallel sentence pairs for translation model training and 33,270,743 monolingual Latvian sentences for the language model training. The system was tuned using the same tuning set and evaluated on the same evaluation set as before. The adapted systems (*Big_Int_LM+T&CC_Terms* and *Big_Int_LM+T&CC_Terms+6th*) were built exactly as the *Int_LM+T&CC_Terms* and *Int_LM+T&CC_Terms+6th* systems from the previous experiment. The results (in Table 6.25) show a relative BLEU increase of 8.8% and 14.9% over the baseline for the system without the 6*th* feature and with the 6*th* feature, respectively. As more data creates higher ambiguity, the 6*th* feature allows increasing the results significantly more than in the previous experiment. This shows the potential of the method when applied on larger corpora.

The results of the experiments show that integration of terminology within SMT systems, even with simple techniques (adding translated term pairs to the parallel data corpus or adding an in-domain language model), can achieve improvement in SMT system quality by up to 23.1% over the baseline system. Transformation of translation model phrase tables into term-aware phrase tables can boost the quality up to 24.1% over the baseline system, mostly because of wrong translation candidate filtering in the translation process.

The experiments also show that the usage of pseudo-parallel sentence pairs extracted from weakly comparable narrow-domain corpora and term pairs acquired from term banks without a sophisticated term sense disambiguation, and semantic analysis of the source text may not result in increased SMT quality due to the added noise in in-domain translation hypotheses.

Table 6.24 English–Latvian automotive domain SMT system adaptation results

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
Int_LM	11.30	10.61	3.9606	3.8190	89.74	90.34	0.1736	0.1312
In-domain_LM_only	11.16	10.52	3.9447	3.8074	89.31	89.92	0.1726	0.1305
Int_LM+T_Terms	12.93	12.12	4.2243	4.0598	88.58	89.32	0.1861	0.1418
Int_LM+T&CC_Terms	13.50	12.65	4.2927	4.1105	88.86	89.70	0.1878	0.1443
Int_LM+ETB_Terms	11.26	10.52	3.9456	3.7882	89.43	90.04	0.1737	0.1290
Int_LM+LEXACC_0.35	10.75	10.09	3.7935	3.6682	90.31	90.86	0.1646	0.1229
Int_LM+LEXACC_0.51	11.08	10.28	3.9132	3.7709	90.23	90.78	0.1706	0.1286
Int_LM+T_Terms+6th	13.19	12.36	4.2657	4.0962	88.84	89.62	0.1876	0.1439
Int_LM+T&CC_Terms+6th	13.61	12.78	4.3514	4.1747	88.54	89.32	0.1920	0.1469

Table 6.25 English–Latvian automotive domain big SMT system adaptation results

System	BLEU	BLEU (CS)	NIST	NIST (CS)	TER	TER (CS)	METEOR	METEOR (CS)
Big_Baseline	15.85	15.00	4.84	4.69	73.80	75.12	0.2098	0.1651
Big_Int_LM +T&CC_Terms	17.24	16.12	5.00	4.83	72.16	73.59	0.2163	0.1717
Big_Int_LM +T&CC_Terms +6th	18.21	17.08	5.15	4.96	70.22	71.62	0.2191	0.1747

6.5 MT Adaptation to a Narrow Domain in Case of Resource-Rich Languages

The objective of this contribution is to evaluate improvements achieved by using data from comparable corpora for tuning Machine Translation systems to narrow domains for languages that are usually classified as resource-rich. The language direction chosen was German to English, and the automotive domain, in particular the sub-domain on transmission/gearbox technology, was selected as an example for a narrow domain. In order to assess the effect of domain adaptation on MT systems with different architecture, both data-driven (SMT) and knowledge-driven (RBMT) systems were evaluated.

6.5.1 Evaluation Objects: Narrow-Domain-Tuned MT Systems

The evaluation objects are two versions of an MT system: a baseline version, without domain tuning, and an adapted version, with domain tuning. Their comparison shows whether or not domain adaptation can improve MT quality.

The evaluation objects were created as follows:

1. For the baseline systems, on the RBMT side, the system of Linguatec's '*Personal Translator*' PT (V.14) was used which is a rule-based MT system based on the IBM slot-filler grammar technology (Aleksić and Thurmair 2011). It was taken as out of the box and installed on a standard PC. On the SMT side, a baseline Moses system with standard parallel data (Europarl, JRC, etc.), which was presented in Sect. 6.2.3, and some initial comparable corpus data as collected in ACCURAT (Skadiņa et al. 2010) were used.
2. For adaptation of the baseline systems, data was collected from the automotive domain. This data was collected by crawling sites of automotive companies that are active in the transmission field (like ZF, BASF, Volkswagen and others). This data was strongly comparable. It was then aligned and cleaned manually. Some sentence pairs were set aside for testing, and the rest were given to the two

systems for domain adaptation as development and test sets. The resulting narrow-domain automotive corpus has about 42,000 sentences for German-to-English.

For the SMT system, domain adaptation was done by adding these sentences to the training and development sets and building a new SMT system.

In the case of rule-based technology, domain adaptation is more complicated as it involves terminology creation which is the main means of adaptation. Therefore, the following steps were taken:

- Creation of a phrase table with GIZA++ and MOSES; for this, the phrase tables of the SMT adapted system were taken; phrase tables of only in-domain data were also built but turned out to be not as efficient as the ones from baseline plus in-domain data.
- Extraction of bilingual terminology candidates from these phrase tables using the P2G (Phrase-Table-to-Glossary) tool; this resulted in a list of about 25,000 term candidates.
- Preparation of these candidates for dictionary import, including creation of part-of-speech and gender annotations, removal of already existing entries, resolution of conflicts in transfers, etc.; the final list of imported entries was about 7100 entries.
- Creation of a special 'automotive' user dictionary which can be added to the system dictionary in cases where texts from the automotive domain are translated.

This procedure is described in detail by Thurmair and Aleksić (2012).

The result of these efforts was four test systems for German-to-English, tuned for the automotive domain with the same adaptation data:

- *SMT-base:* DFKI-baseline system trained with only baseline data
- *SMT-adapted:* DFKI-adapted system trained with baseline plus in-domain data
- *RBMT-base:* PT-baseline as the out-of-the-box RBMT system
- *RBMT-adapted:* trained with an additional 'automotive' dictionary.

6.5.2 Evaluation Data

For evaluation, a set of sentence pairs was extracted from the collected strongly comparable automotive corpora. In total, about 1500 sentences were taken for tests, with one reference translation each.

The sentences represent 'real-life' data; they were not cleaned or corrected, just like the training data. So they contain spelling mistakes, segmentation errors and other types of noise. This fact, of course, affects the translation quality for the adapted systems.

6.5.3 *Evaluation Methodology*

Several methods can be applied for the evaluation of MT results. **Automatic** comparison (called BLEU in Fig. 6.8) is the predominant paradigm in the world of SMT. So BLEU (Papineni et al. 2002) and/or NIST (Doddington 2002) scores can be computed for different versions of MT system output.

While such scores seem to measure inner-system quality changes with some degree of reliability, they do not seem to measure translation quality (Babych and Hartley 2008), do not conform to the judgment of human evaluators (Hamon et al. 2006), and are sensitive towards an SMT system architecture in disfavour of rule-based approaches. Therefore, projects like WMT do not use them as the only measure of quality any more (Callison-Burch et al. 2009; Bojar et al. 2018) but also ask for human judgment.

Comparative evaluation (called COMP in Fig. 6.8) is possible between two systems as well as between two versions of the same system. It simply asks whether or not one translation is better/equal/worse than the other.

While this approach can find which of two systems has an overall better score, it cannot answer the question of what the real quality of the two systems is: '*Equal'* can mean that both sentences are perfect or that both are unusable.

Therefore, **absolute** evaluation (called ABS in Fig. 6.8) is required to determine the quality of a given translation. This procedure looks at one translation of a source sentence at a time and determines its accuracy (how much content has been transported to the target language) and fluency (how correct/grammatical is the produced target sentence), following the FEMTI paradigm (King et al. 2003).

Post-editing evaluation (called POST in Fig. 6.8) reflects the task-oriented aspect of evaluation (Popescu-Belis 2008). It measures the distance of an MT output to a human (MT-post-edited) output, either in terms of time (answering the question of how productive a system can be as compared, e.g. to a human-only translation) or in terms of the keystrokes needed to produce a human-corrected translation from an MT-raw translation (HTER: Snover et al. 2006, 2009).

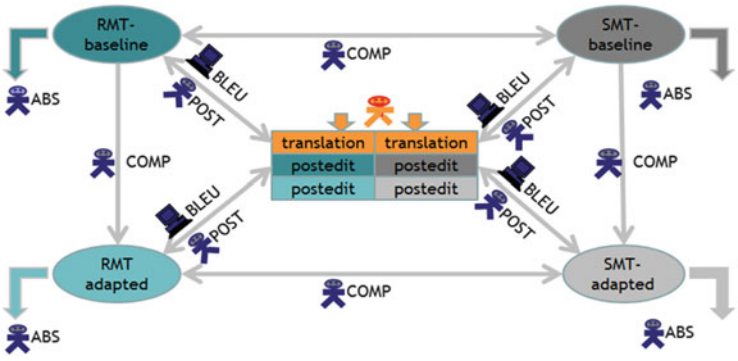

Fig. 6.8 Evaluation options

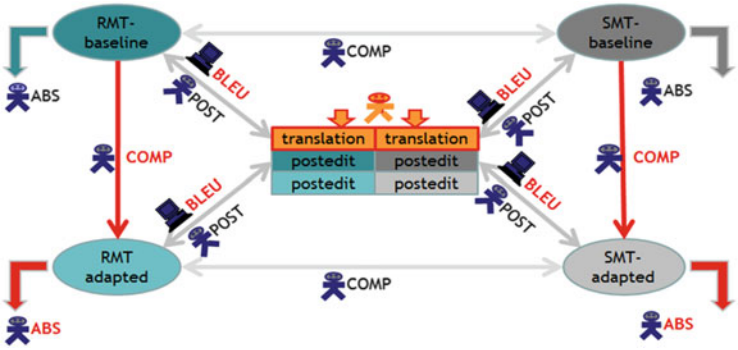

Fig. 6.9 Evaluation in narrow domain task

Post-editing evaluation adds reference translations to the evaluation process.

In our narrow domain task, the following evaluation methods were used, cf. Fig. 6.9:

- Automatic evaluation of the four systems (SMT-baseline and SMT-adapted, RBMT-baseline and RBMT-adapted) using BLEU and NIST scores.
- Comparative evaluation of the pairs (SMT-baseline versus SMT-adapted and RBMT-baseline versus RBMT-adapted); this would produce the core information of how much the systems can improve.
- Absolute evaluation of the systems (SMT-adapted and RBMT-adapted), to gain insight into the translation quality and, consequently, the potential acceptance of such systems for real-world use.

Other forms of evaluation were not included in the evaluation task. However, to have a complete picture, the other ABS and COMP directions were evaluated as well, but with less effort (1 tester only).

6.5.4 Evaluation Tools

To perform the evaluations, a special toolset was created for the non-automatic tasks. The toolset is called '*Sisyphos-II*' (for details see Chap. 8: Appendix) and consists of three components:

- 'ABS' to support absolute evaluation, using two four-point scales. For adequacy, the options are {*full content conveyed* | *major content conveyed* | *some parts conveyed* | *incomprehensible*}. For fluency, the options are {*grammatical* | *mainly fluent* | *mainly nonfluent* | *rubble*}.
- 'COMP' to support comparative evaluation of two MT outputs, using a four-point scale. Comparison options are {*first translation better* | *both equally good* | *both equally bad* | *second translation better*}.

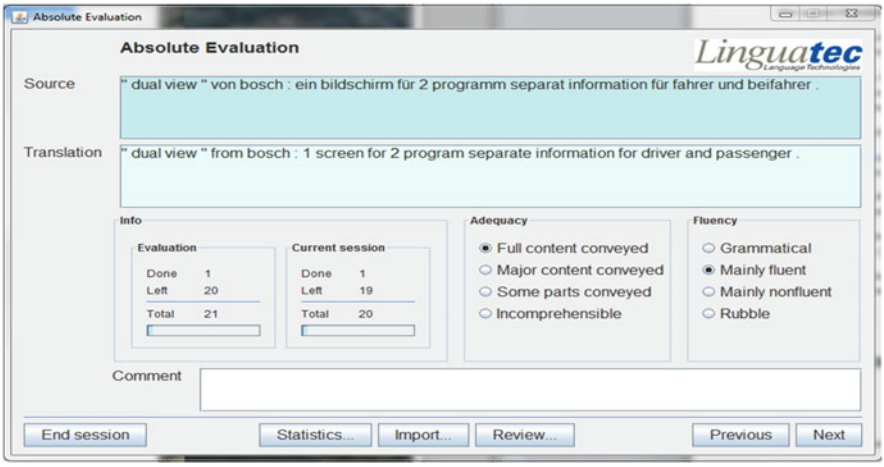

Fig. 6.10 Screen shot of evaluation tool (ABS)

- 'POST' to support post-editing evaluation, by measuring the post-editing time from the first display of the sentence until the pressing of the [Save] button (in seconds) and allowing HTER computing.

The tools are stand-alone tools that can be given, for example, to a freelance translator. Evaluation data is presented to the users by a special GUI in random order, and evaluation results are collected in an XML file which is the basis for evaluation.

An example screenshot of the tool is shown in Fig. 6.10. Each time a 4-point scale is presented, users select one of the options in both areas.

6.5.5 Evaluation Results

Three evaluators were used to do the translations, all of them good speakers of English with a bit of MT background. Each of them evaluated a random subset of the 1500 sentence test set, consisting of at least 500 sentences for each of the COMP evaluations (SMT-adapted versus SMT-baseline and RBMT-adapted versus RBMT-baseline) and at least 300 sentences for each ABS evaluation (SMT-adapted and RBMT-adapted). More than 5000 evaluation points were collected this way.

6.5.5.1 Automatic Evaluation

The automatic evaluation for the German–English pair was done on the basis of BLEU scores. The results are shown in Table 6.26.

Table 6.26 BLEU scores for SMT and RBMT

	SMT	RBMT
Baseline	17.36	16.08
Adapted	22.21	17.51
Improvement	**4.85**	**1.43**

For both systems, there is an increase in BLEU; it is more moderate for the RBMT than for the SMT system. However, it is known that BLEU is biased towards SMT systems.

6.5.5.2 Comparative Evaluation

For the German–English pair, three testers were used, all of them good speakers of English with a bit of MT background.

Of the 1500 test sentences, three testers inspected randomly selected subsets, in total about 2000 sentences. As the tool does not offer identical sentences for evaluation, these cannot be differentiated for 'equally good' versus 'equally bad'. If these two categories are merged into one ('*equal*'), the following results were achieved (Table 6.27).

The data shows that the domain adaptation results in an improvement of 5% for both types of systems. It is a bit more (5.1%) for the SMT than for the RBMT (4.7%). The result is consistent among the testers: all of them state an improvement of the adapted versions, and all of them see a higher improvement for the SMT than for the RBMT.

It may be worthwhile to notice that in the RBMT evaluation, a large proportion of the test sentences (nearly 60%) came out identical in both versions, and the changes were rather small (17% of the sentences). In the SMT system, nearly no sentence came out unchanged, and the variance in comparison was between 36% and 51% (depending on the testers).

In a sideline evaluation, a comparison was made between the baseline versions of SMT and RBMT and their adapted versions (Table 6.28).

The result shows that the RBMT quality is considered significantly better than the SMT quality. The main reason for this seems to be that the SMT German–English frequently eliminates verbs from sentences: for example *Silber wird in der Medizin seit Jahrhunderten wegen seiner antimikrobiellen Wirkung geschätzt und eingesetzt. => silver in medicine centuries for its antimicrobial effect and*. This effect has already been observed with other SMT outputs.

It should be noted, however, that the distance between the systems is smaller in the adapted versions than in the baseline versions (by 3%).

6.5.5.3 Absolute Evaluation

The absolute evaluation was done to assess how usable the resulting translation would be after the system was adapted. A total of 1100 sentences, randomly selected from the 1500 test base, were inspected by three testers. The adequacy and fluency

Table 6.27 Comparative evaluation baseline versus adapted for SMT and RBMT

	SMT					RMT				
	Total inspected	Base better	Both equal	Adapted better	Improvement (%)	Total inspected	Base better	Both equal	Adapted better	Improvement (%)
Tester 1	1049	235	514	300	6.20	1501	91	1237	173	5.46
Tester 2	510	130	228	152	4.31	503	33	417	53	3.98
Tester 3	501	82	319	100	3.59	501	34	418	49	2.99
Total	2060	447	1061	552	**5.10**	2505	158	2072	275	**4.67**

Table 6.28 Comparative Evaluation SMT / RBMT, baseline and adapted

Total inspected	SMT better	Both equal	RBMT better	In percent
501	47	170	284	47.3
489	38	203	260	44.3

was measured for each sentence on a scale of 1–4. Table 6.29 gives the result (lower average scores mean better quality).

It can be seen that testers evaluate the SMT somewhat between 'mainly' and 'partially' fluent/comprehensible and the RBMT close to 'mainly' fluent/comprehensible. If the percentage level of 1/2 of the evaluations is taken, the SMT adequacy is rated with 36.6% and fluency with 53.04%, while both adequacy (64.97%) and fluency (77.50%) are significantly higher in RBMT. All testers agree in their evaluation and have similar average results. The better score for RBMT may result from the 'missing verb' problem mentioned above.

It could be worthwhile to mention that the often-heard opinion that SMT produces more fluent output than RBMT cannot be corroborated with the evaluation data here: the RBMT output is clearly considered to be more fluent than the SMT output (1.8 vs. 2.3).

An absolute evaluation was also done for the two baseline systems, however with only one tester. The results are given in Table 6.30.

The figures indicate that system adaptation improves the accuracy of both of the SMT (from 2.86 baseline to 2.62 adapted), and it seems to reduce the fluency of the RBMT (from 1.48 baseline to 1.80 adapted); a further error analysis would be required to find out why. The other results (RBMT accuracy and SMT fluency) seem unchanged.

As far as the inter-rater agreement is concerned, the test set-up made it difficult to compute it: all testers used the same test set but tested only a random subset of it. So there are only a few data points common to all testers (only 20 in many cases). For those, only weak agreement could be found (with values below 0.4 in Cohen's kappa, Table 6.31). However, all testers show consistent behaviour in the evaluation and came to similar overall conclusions, as has been explained above.

6.5.6 Conclusion

Figure 6.11 gives all evaluation results. The main conclusion is that all evaluation methods indicate an improvement of the adapted versions over the baseline versions.

Automatic evaluation:

- For SMT, the BLEU score increases from 17.36 to 22.21.
- For RBMT, the BLEU score increases from 16.08 to 17.51.

Comparative evaluation:

- For SMT, an improvement of 5.1% was found.
- For RBMT, an improvement of 4.67% was found.

Table 6.29 Absolute evaluation for SMT-adapted and RBMT-adapted systems

	Inspected	Adequacy						Fluency					
		1: full	2: most	3: partial	4: none	Average	% of 1+2	1: fluent	2: mainly	3: partly	4: none	Average	% of 1+2
		SMT adapted											
Tester 1	500	89	119	284	8	2.42	41,60	87	163	238	12	2.35	50,00
Tester 2	302	52	48	156	46	2.65	33,11	97	97	93	15	2.09	64,24
Tester 3	301	59	37	77	128	2.91	31,89	116	25	31	129	2.57	46,84
Total	1103	200	204	517	182	2.62	**36,63**	300	285	362	156	2.34	**53,04**
		RMT adapted											
Tester 1	501	210	127	150	14	1.94	67,27	197	189	100	15	1.87	77,05
Tester 2	300	106	99	80	15	2.01	68,33	164	89	42	5	1.63	84,33
Tester 3	301	149	25	55	72	2.17	57,81	180	35	34	52	1.86	71,43
Total	1102	465	251	285	101	2.02	**64,97**	541	313	176	72	1.80	**77,50**

Table 6.30 Absolute evaluation of the baseline systems

	Inspected	Adequacy						Fluency					
		1: full	2: most	3: partial	4: none	average	% of 1+2	1: fluent	2: mainly	3: partly	4: none	average	% of 1+2
SMT-baseline	301	57	51	69	124	2.86	**35,88**	136	22	46	97	2.35	**52,49**
RMT baseline	301	165	15	61	60	2.05	**59,80**	222	37	18	24	1.48	**86,05**

Table 6.31 Kappa for inter-tester agreement

	SMT COMP	RMT COMP	SMT-ABS adequacy	SMT-ABS fluency	RMT-ABS adequacy	RMT-ABS fluency
Records inspected	1189	1102	846	846	851	851
Common data points	115	39	21	21	21	21
Common evaluation	46	11	5	4	3	3
Kappa	0.38	0.26	0.22	0.18	0.17	0.11

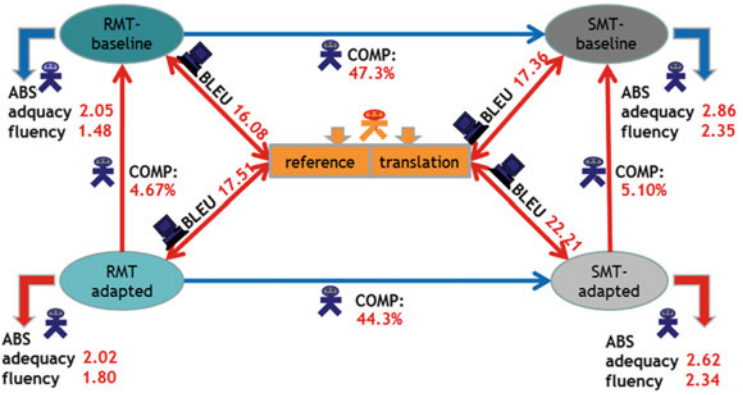

Fig. 6.11 Evaluation summary (BLEU, COMP, ABS)

Absolute evaluation:

- For SMT, adequacy improved from 2.86 to 2.62, and fluency improved slightly from 2.35 to 2.34.
- For RBMT, adequacy improved from 2.05 to 2.02, and only fluency decreased from 1.48 to 1.8.

The improvement is more significant for the SMT system than for the RBMT. This may be due to the fact that the RBMT baseline system has better COMP and ABS scores, though lower BLEU scores, than the SMT baseline.

For SMT improvement, Pecina et al. (2012) report improvements between 8.6 and 16.8 BLEU (relative) for domain adaptation. Our results here are in line with these findings.

6.6 Application of Machine Translation (MT) in Web Authoring

Authoring is defined as a process of creating and editing documents, especially multimedia documents, for other end-users. Authoring systems or tools are software packages used for creating and packaging content and have been applied in multimedia (Bulterman and Hardman 2005; Scherp and Boll 2005; Deltour and Roisin 2006), e-learning (Watson et al. 2010; Capuano et al. 2009), adaptive e-learning (Bontchev and Vassileva 2009), mobile learning (Mugwanya and Marsden 2010), tutoring (Escudero and Fuentes 2010), interactive digital storytelling (Müller et al. 2010) and lately digital gaming (Mehm et al. 2012). Their purpose is to assist less technically skilled users to produce multimedia, structure the content without special expertise and speed up the process of content creation by streamlining and automating common tasks.

Authoring tools are also widely used in professional publishing for a variety of publication types such as books, documentations, reports, articles or presentations. The simplest authoring tools are simple preformatted Word document templates or templates with added scripting. More sophisticated publishing systems, platforms and desktop publishing applications enable advanced functionalities and may support different roles in the system, for example writers and editors, and even provide the ability to collaborate between these roles. Web authoring is a sub-domain of authoring and, in its broadest sense, means authoring content online.

The way people use the Web has changed considerably. Web users are provided with new means of communication (blogs, tweets, instant messaging), collaboration (wiki, forums), and sharing of multimedia content. From the late 1990s, user-oriented Web authoring started changing the publishing game. Previously, special technical knowledge was required to create and publish content on the Web, but the emergence of Web authoring tools has brought publishing closer to the wider public. Web authoring has facilitated content creation by non-technical users: all a user needs is an Internet connection and a browser, everything else is available online. Furthermore, Web authoring tools started providing simple interfaces that guide and assist at every step of content creation by enabling users to develop websites in desktop publishing (or similar) format by generating underlying HTML code for the layout based on the user's design. Users can typically toggle between graphical design and HTML code.

Web authoring systems that are accountable for most user-generated content are popular blogging platforms (such as WordPress[5] and Blogger[6]), micro-blogging platforms (such as Twitter[7] and Tumblr[8]) and wiki platforms (Désilets et al. 2006)

[5]WordPress: http:www.wordpress.com

[6]Blogger: http://www.blogger.com/

[7]Twitter: http://twitter.com

[8]Tumblr: http://www.tumblr.com

based on the MediaWiki[9] open source project. We consider social media networks as a part of Web authoring because users can create multimedia profile pages, post text, images, video or a combination of all of them and create content in this way.

The availability and popularity of Web authoring tools have affected many areas, including translation of online content. The amount of data has expanded drastically, the number of languages in which online content is produced has increased, and even English content is frequently written in *casual* English. Use of Web authoring (and the Web in general) has shifted from producers to users who connect and exchange ideas and opinions, but the language barrier still remains. Garcia (2009) stated that *the amount of content contributed by producers and users exceeds translation industry's capacity to cope* and that the translation industry *cannot keep pace with an environment that puts a premium on cheapness and speed*.

Today, researchers working in the field of natural language processing, specifically information retrieval and machine translation (MT), are faced with the seemingly low-hanging fruit of the large amounts of online data that is provided through user-oriented media (social networking sites, blogging and microblogging services). In reality, this comes with a price, bringing new issues they have not faced before.

To illustrate the vastness of online content, we provide figures published by the most popular services: Tumblr claims to have 277 million blogs in 16 languages and 128.7 billion posts[10]; bloggers at WordPress.com write in 120 languages, publish 41.7 million new posts and 60.5 million new comments per month, and 409 million people view more than 15.5 billion pages each month; Twitter supports more than 35 languages and has more than 320 million active users[11]; Wikipedia has 5 million content articles in English only, is available in 291 languages, and more than 19,000 articles are added to it every day.[12] These numbers give us a perspective on the amount of content that users generate. Researchers have to deal with scaling-up demands and the robustness required by the need to understand *casual written English, which often does not conform to rules of spelling, grammar and punctuation* (Clark and Araki 2011). Online content is ever more resembling conversation with loose grammar rules, intentional or unintentional misspellings, acronyms and jargon which affects the accuracy of natural language processing, information retrieval and translation.

6.6.1 The Role of Translation and MT in Web Authoring

Web authoring is a multi-language online environment. With the aid of Web authoring, we now have the means to search for information globally or locally,

[9]MediaWiki: http://www.mediawiki.org/

[10]https://www.tumblr.com/about, accessed in January, 2016.

[11]https://about.twitter.com/company, all numbers approximate as of September 30, 2015.

[12]http://stats.wikimedia.org/EN/ReportCardTopWikis.htm, accessed in January, 2016.

while social media tools help us to distribute it to various communities that speak other languages and, in this way, to broaden readership and/or the pool of customers. Social tools also allow us to build and support international communities or networks communicating in languages other than English or our mother tongue.

Yet there is still one important obstacle in the way of reaching the full potential: the lack of quality translation, especially for under-resourced languages and narrow domains.

The role and the importance of translation in Web authoring can be viewed from different perspectives, and we will focus on two that are closely related to Web authoring, namely

- Content creator perspective, which includes companies or organisations publishing content and public publishing user-generated content via social media tools and platforms (blogging, microblogging, wikis, forums).
- Content consumer perspective which involves readers searching for information or opinions.

From the content creator perspective, we distinguish between localisation and internationalisation. In this context, localisation is considered as a translation from English (or any other major world language) into other languages, including under-resourced languages and minority languages; for internationalisation, the direction of translation is exactly the opposite: translating from a (minor) language into one of the major languages. While localisation is more typical for larger companies, internationalisation is more frequent for smaller companies or bloggers wanting a larger audience. For example large international companies are expanding their business to other countries and want to localise their Web pages, product documentation or user support. On the other hand, small (local) companies want to reach out and provide their content, especially Web sites, in one of the more frequently used languages.

From the end-user perspective, translation plays an important role in discovering new knowledge, finding information about products, people and events. Translation direction can range from English to any language, between any language pair, even from an under-resourced one to another under-resourced one.

Quality is an important factor in the translation in Web authoring but is not the only one. When readers are just interested in the broader meaning of the text (so-called *gisting*), the quality is less important than the speed of translation and its accessibility to the public.

The traditional translation model includes the aid of computer translation software and is carried out by professional translators or bi-lingual experts. Most translation software was based on translation memories and terminology databases, thus being less suitable for the needs of translation in Web authoring. The translation model has changed with the rise of ‘software as a service’. Translation services, particularly the ones based on MT, are now more affordable, available to the general public, and suitable for integration into authoring tools, more easily than ever before.

From the content consumer perspective, MT (as a service) seems to be the only good option for translating user-generated content. In general, readers do not

understand more than two or three languages, cannot afford a human translator and do not want to buy expensive translation software. Content producers see MT as important for similar reasons; they want translation to be as fast and as cheap as possible.

Translation quality is a major factor in Web authoring, but it is not the most important factor in some cases. When readers are just interested in the broader meaning of the text (so-called gisting), the quality is less important than the speed of translation, public availability and price—especially if the service is free.

The reasons observed by Hutchins (2003) regarding why machine translation is needed are still valid today, also for the domain of Web authoring. We added the lack of support for under-resourced languages to the following list of reasons:

- The amount of generated content is too large for human translators.
- The demand for increase in the volume and speed of translation throughput (translation needed now, not in a few days from now) is growing.
- Top quality translation is not always needed, neither is human assistance/post-editing.
- People communicate and generate content in a large number of under-resourced languages, which are usually not supported by traditional translation models or are not easily accessible.

An additional reason might also be the need for integration of translation service into other tools used for research, exploration and discovery. For example, corporations that use multi-lingual online collaboration environments need translation tools that seamlessly integrate into collaborative tools, such as chats and support forums, for more effective use of online content (replacing the need to copy-paste content into one of the freely available MT services) and appealing to an even broader population of users.

While translation services are valuable standalone products, they are more valuable if they can be integrated to complement the functionalities in other tools that users work with, such as Web browsers, document editors, phone applications and Web authoring tools and platforms.

6.6.2 Characteristics and Requirements for Translation in Web Authoring

In terms of demographic factors (such as geographic location, age, gender, household income or levels of education), Web authoring is widespread mostly due to the emergence of new online tools and platforms that make Web authoring easier than ever before. Translation in Web authoring is needed, and it has to meet the requirements set by its users. Web users are very demanding—they want translation to be fast and free.

The role of human translators in Web translation has changed: the old traditional translation model of translate-edit-proofread, involving human (professional) translators, has been replaced with other, more flexible models—not collaborative models, but rather MT-assisted models (Garcia 2009).

A number of MT-assisted models are already implemented in translation services, such as Google Translate or Microsoft Translator, and are being widely used. However, before using them, we have to consider the following question: is MT really the answer to every translation problem in Web authoring? Considering the current state of MT, the answer to this question is not encouraging. Several factors have to be considered before deciding to use MT:

- Role of the user: content consumer or content producer.
- Volume of material: the larger the volume, the more prohibitive the cost of human translation becomes.
- Frequency with which material changes: it may be less practical to continually use human translators for material that changes frequently.
- Domain and purpose of content and its translation: informational, persuasive, legal, etc. The more important it is that the translation is accurate and fluent, the less likely it is that MT should play a role, at least not without post-editing.
- Speed of translation: MT will always provide faster results.
- Languages involved: related languages and languages that are very commonly used will translate the best. For some language pairs, it might be hard to find a human translator and be much easier to use Google Translate, even if the translation quality is not good.

Balancing these factors is an important part in making the decision of whether to apply MT or not. The translation quality of MT tools depends on the domain and the languages involved, and therefore, it is important to choose the tool that is best suited for the problem, as not all tools produce good results with all language pairs.

Current translation techniques that are applied to Web authoring depend on the type of platform and the content. Popular content management systems (CMS) allow editing of multi-lingual content in parallel at the time of writing or soon afterwards. In most cases, authors translate text themselves on the fly. For Web authoring platforms, such as WordPress, several plug-ins provide the functionality of parallel text editing of multiple languages.

The collaborative translation model is used for wiki projects such as Wikipedia and Wikitravel. They both provide content in multiple languages, and translation is performed by multiple (anonymous) bi-lingual authors/editors. However, some might not truly consider this to be multi-lingual translation, because neither Wikipedia nor Wikitravel provides (exactly) the same content in different languages. Research by Désilets et al. (2006) has refuted the commonly held assumption that Wikipedia contents are parallel, they claim that '*[t]hese sites are in fact a collection of parallel communities that produce content about overlapping sets of topics in different languages, with little if any synergy across languages*'. There is also a

project translatewiki.net[13] which is a wiki localisation platform for translation communities, language communities and free and open source projects. The platform incorporates translation memory from the translate toolkit,[14] Yandex Translate[15] and Microsoft Translator which assist in collaborative translations.

Crowdsourcing translation is a similar model to the collaborative one, but it is not limited to the wiki environment. Facebook crowdsourced its translation, and according to the results, this could not be performed any faster or better even if it had applied the usual localisation processes (Garcia 2009). Twitter is using a similar approach by inviting its users to help with localisation of the platform. Some of the other 'big players' have also used crowdsourcing to translate the parts of their content that they considered to be suitable for this kind of translation. For example Google used crowdsourcing to translate its interface into many minority languages. It also uses the '*Suggest a better translation*' feature in Google Translate through which crowds contribute to improvements of its SMT engine (Garcia 2009).

6.6.2.1 MT in Web Authoring

Traditional translation models involving professional translators and the workflow of using only translation memories are less suitable for fast growing (in terms of the volume of publications) and expanding (in terms of new languages) Web authoring domain, especially if compared with MT systems.

Today's widely spread use of MT on the reader side of Web authoring can be credited mostly to Google Translate and Microsoft Bing Translator. They are typically used at the time when content is consumed. Readers have the option of using a Web browser with an integrated translation service available via a toolbar or installation of the tool as an extension for their favourite Web browser. When readers visit a certain website, if the content language differs from the default language set in the browser, it is either automatically translated or translated on demand by pressing a button on the toolbar. Some bloggers put special translation widgets directly on their blogs, so readers do not have to install toolbars and can use the translation widget instead.

The situation is similar for content creators. For example bloggers can use several translation plug-ins which are available for the most popular blogging platforms. These plug-ins usually use Google Translate or Bing Translator to translate text in the Web editor, and they put it directly back in the editor so that the author can post-edit it before publishing. The microblogging platform Twitter has integrated Bing Translator API into their Web–user interface to provide machine translation between more than 40 language pairs.

[13]Translatewiki project: http://translatewiki.net/wiki/

[14]Translate Toolkit & Pootle: http://translate.sourceforge.net/wiki/

[15]Yandex Translate: http://company.yandex.com/technologies/translation.xml

Wiki projects are a special case in regard to machine translation. Wikipedia took the initiative in the form of the Wikipedia Machine Translation Project.[16] As Wikipedia is a multi-lingual resource, the 'Wikipedia consensus is that an unedited machine translation, left as a Wikipedia article, is worse than nothing.'

6.6.2.2 Translating User-Generated Content

Web authoring covers online content by professionals and amateurs. The latter is also known as user-generated content. It is usually produced in a more conversational manner, most of it is in poor or non-standard quality, it can be produced by non-native speakers, native speakers can non-deliberately introduce typos or deliberately stray from spelling norms to achieve special linguistic goals or effects (Jiang et al. 2012).

Carrera et al. (2009) acknowledged that user-generated content is suitable for MT, but most such content usually remains untranslated. Jiang et al. (2012) built a number of statistical SMT engines for a Middle East–based social networking provider based on user-generated content and identified several problems in the process.

Flournoy and Rueppel (2010) describe how MT could be used in Adobe for translating user-generated content either for a community translation initiative, in which MT output can be presented as pre-translations for the members of the community, or for translating valuable resources such as Q&A, tutorials and product reviews. While high-quality MT is preferred in both cases, it is not required.

Evaluation of MT is a separate research field, and we will not delve far into it. In many studies including the one by Hovy et al. (2002), the following aspects of translation quality are taken into consideration: fluency (lexically and syntactically well-formed sentences), fidelity (translation does not change the meaning/semantics of the input), price, system extensibility and coverage (specialisation of the system to the domains of interest). More recent research studies about translating user-generated content were mainly interested in fidelity. Fidelity is measured on a limited scale by human judges rating how well a system's output expresses the content of the same portion of the source text or even ideal human translations (Hovy et al. 2002). Mitchell and Roturier (2012) conducted a pilot study, based on a previous study by Roturier and Bensadoun (2011), that examined the perceived quality of MT in terms of comprehensibility among members of an online community forum and the ways users interact with the MT content. Even though the study had a low response rate, the results have shown that the MT output was *comprehensible slightly more often than not*.

Translation direction in user-generated content is primarily from English to other languages; otherwise, it varies and can include any language pair. Open-source MT

[16] https://meta.wikimedia.org/wiki/Machine_translation

translation attempts are an opportunity for minor languages, and the objective behind is also to 'de-minorise' translation (Forcada 2006).

From the usage of MT in Web authoring, we can conclude that it is mostly used and useful for obtaining a general understanding of content. If it is used for content creation, then the content is post-edited, because the quality of MT is not good enough.

6.6.2.3 Defining Requirements for Using MT in Web Authoring

When defining requirements for MT in Web authoring, we have to consider both content characteristics and the factors that we mentioned at the beginning of this section. Major factors affecting the quality of MT in Web authoring are

- **Domain specificity:** MT systems based on texts in one domain perform badly in another domain; may work well for general translations but not for specific ones, or vice versa; work well for EU-related documents, but perform really bad for general translations.
- **Lack of resources**: SMT systems rely solely on quantitative information extracted by systems trained on vast amounts of data. What if there are no vast amounts of data for systems to be trained on, as in cases of under-resourced languages or narrow domains?
- **Casual English**: problems include rapidly changing out-of-dictionary slang, short-forms and acronyms, punctuation errors or omissions, phonetic spelling, misspelling for verbal effect and other intentional misspelling and recognition of out-of-dictionary named entities. Use of casual English in social media poses a problem: casual media needs pre-processing before translation, but this might not prove to be feasible for bloggers (Clark and Araki 2011).

Requirements for MT on translating cross-language social media were described by Carrera et al. (2009) in the context of social media analysis. They noted that an MT system would need to be designed for

- Large-scale, real-time translation
- Preservation of meaning (which should be good enough for gisting)
- Robustness, especially in light of errors in linguistic formalisation

Flournoy and Rueppel (2010) provide additional requirements valid for translating user-generated content:

- Low to medium translation quality is required.
- MT has to be able to deal with various subject matters.
- There is no need for special security (no need for non-disclosure agreements as in the case of formal documents with business secrets).
- The most frequently occurring language pairs are EN→XX, but others can also occur, such as XX→YY.
- Input is of varied, uncontrolled quality.

Almost all researchers agree on the biggest issue that all MT systems are facing: the quality of translation output. If we ignore the fact that most MT systems prior to Google Translate were either rule-based or assisted by translation memories, one of the more important causes for poor quality is the discrepancy between the corpora that MT systems are trained on and the texts that MTs are used on. For example, Google Translate works best for short subject–verb–object sentences, such as driving directions, simple instructions or simple scientific sentences. It also does quite well for gisting of websites, but is unlikely to provide adequate translations for short-lived colloquialisms, new words or word plays. Using Google Translate to directly translate social media content without post-editing is not recommended. The same goes for legal drafts, descriptions of medical equipment, political texts, safety applications and legal documents.

6.6.3 MT Systems Enhanced with Comparable Corpora in Web Authoring: A Use Case

The quality of translation services for under-resourced languages and narrow domains still falls behind the quality for more widely used language pairs (e.g. English, German, French, Arabic, Chinese) and more general domains. MT systems enhanced with comparable corpora aim to close this gap and improve the quality of translation for these under-resourced languages and narrow domains.

Comparable corpora are easier to obtain than parallel corpora, but, in comparison to other comparable corpora in major languages, they are still not in abundance. Content from narrow domains faces a similar situation: translation services trained on general texts produce poor results when used on texts from narrow domains and make it hard to train a quality SMT due to the lack of parallel corpora.

The blogosphere is a good example that combines both of the above-mentioned issues, which were addressed in the ACCURAT project. We evaluated the use of MT systems enhanced with comparable corpora (henceforth CC-enhanced MT) for Web authoring in a use case involving blog posts in Slovenian, Croatian and German. CC-enhanced MT was used as an intermediate step between content written in one of the under-resourced source languages and Zemanta's recommendation engine that is available via a Web service. Currently, the recommendation service works only for texts in English and does not return good results for texts in other languages, so this was an opportunity to use translation before the recommendation step returns more relevant related contents.

6.6.3.1 Evaluation Process and Datasets

As a blogger writes a blog post, Zemanta's recommendation engine analyses the text and suggests related contents that the blogger can use to enrich the blog post. Our goal was to find out whether using MT before sending text to the recommendation

engine results in better suggested related articles. Although the recommendation engine returns related articles, images and keywords, we focussed on related articles only.

Evaluation of results was done in Zemanta's internal evaluation system by two human evaluators. We collected blog posts and online news articles in Slovenian, Croatian and German, 100 texts per language (Table 6.32), and put them through the recommendation engine to obtain the 10 best suggested related articles per text. Texts in the source language were translated using two translation methods—baseline and CC-enhanced MT—and translations were sent to the recommendation engine to get suggestions again.

The evaluation cycle is illustrated in Fig. 6.12. After the recommendation engine returned suggested related articles, two human evaluators assessed each suggested article, from the blogger's perspective, and assigned a score to it ranging from zero (will not use) to three (definitely will use). These scores were used to calculate the precision@10 metric which considers only the top ten relevant documents with the highest precision score.

Table 6.32 Evaluation sets of texts

Language pair	Number of files	Avg. text length (words)
Slovenian–English	100	238.8
German–English	100	242.7
Croatian–English	100	202.7

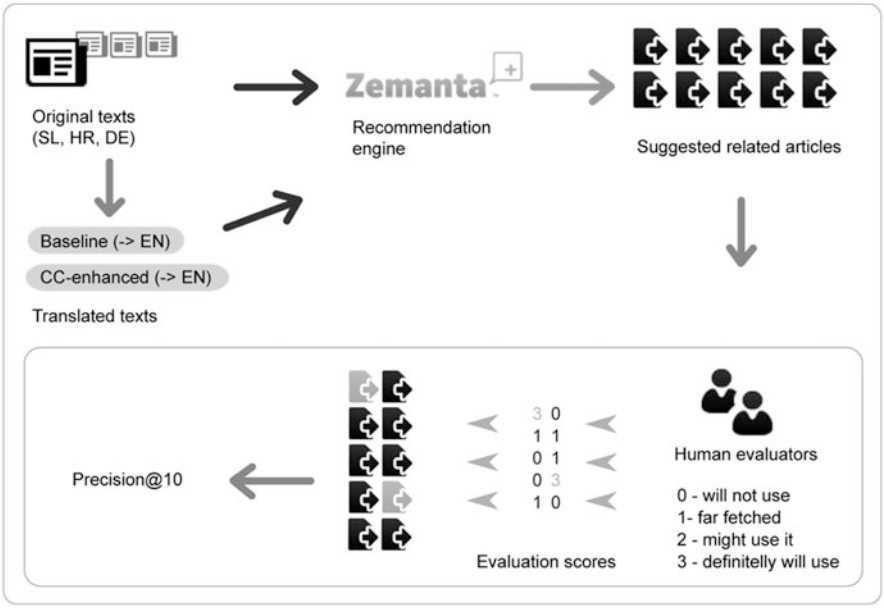

Fig. 6.12 Evaluation process used in the use case

The length of blog posts and online news articles can vary from a few sentences to full-length and detailed reviews. Datasets contain texts of length between 200 and 300 words. This is a typical length of a blog post. This amount of text is enough for the recommendation engine to return valid suggestions and for a translation service to provide translations in a reasonable amount of time. Texts included topics from business/economy, politics, technology, sports and living.

6.6.3.2 Results and Discussion

We calculated precision@10 for all language pairs in three different batches and summarised the average precision in the tables below. Batches are labelled *Original*, *Baseline* and *CC-enhanced*. In the first batch (*Original*), untranslated texts were fed directly into the recommendation engine; in the second batch (*Baseline*), texts were first translated using baseline MT and then fed into the recommendation engine. In the last and third batch (*CC-enhanced*), texts were translated into English using CC-enhanced SMT and then again fed into the recommendation engine. The inter-rater agreement for two human evaluators was moderate which was rather expected due to the high level of subjectivity in the blogosphere.

Table 6.33 shows the average precision for different language pairs. For the Slovenian–English language pair, usage of baseline MT in comparison to original texts improved precision by 11%, and usage of CC-enhanced SMT improved it by 15%.

We can also see that using translation for German texts shows even greater improvement: nearly 20% for baseline MT and 24% for CC-enhanced SMT in comparison to original texts.

Unfortunately, we were not able to use CC-enhanced SMT for Croatian texts, and therefore, we only have precision for baseline MT. Using this translation method improved results by 11%.

We tested the hypothesis that results obtained by using the recommendation engine on MT translated texts does not differ significantly from results obtained by using the engine on untranslated (original) texts using the unpaired t-test. We tested both translation methods, and the difference for all translation pairs was significant on 95% confidence level. Table 6.34 contains mean, SD and P-values for both translation methods. Although the CC-enhanced method improved precision for Slovenian and German, the difference between both translation methods is not significant.

Table 6.33 Average precision for different language pairs

Dataset	Slovenian–English	German–English	Croatian–English
Original	0.153	0.141	0.201
Baseline	0.265	0.344	0.314
CC-enhanced	0.299	0.379	–

Table 6.34 Mean, SD and P-value for language pairs and translation methods

Value	Baseline			CC-enhanced	
	Slovenian	German	Croatian	Slovenian	German
Mean	0.265	0.344	0.313	0.301	0.381
SD	0.245	0.248	0.312	0.243	0.298
P-value	0.0006	0.0001	0.0110	0.000	0.000

Table 6.35 Average, minimum and maximum translation times for baseline method and CC-enhanced method

	Avg translation time (s)		Min time (s)		Max time(s)	
Language pair	Baseline	CC-enhanced	Baseline	CC-enhanced	Baseline	CC-enhanced
Slovenian–English	111.98	133.99	61.56	30.95	365.05	423.88
German–English	172.71	186.98	92.16	122.06	273.62	304.94
Croatian–English	78.92	–	31.70	–	122.47	–

Table 6.36 Percentage of translated words for baseline translation method and CC-enhanced method

	Avg. words			% translated words	
Language pair	Original	Baseline	CC-enhanced	Baseline	CC-enhanced
Slovenian–English	238.8	232.0	225.2	59	76
German–English	242.7	209.8	217.1	73	74
Croatian–English	202.7	183.8	–	73	–

Although we were more concerned with the criteria of fidelity, we also measured the translation time for whole datasets (Table 6.35) and the percentage of translated words for 10 randomly translated texts per translation method and language pair (if available), as summarised in Table 6.36. While average translation times were roughly similar for Slovenian and German, the difference for minimum and maximum time was quite large, but we did not investigate it any further at this point.

Next, we analysed ten randomly selected translated texts per translation method and a language pair (if available) for percentage of translated words. Results are summarised in Table 6.36.

Interestingly, the percentage of translated words when using the CC-enhanced method for Slovenian texts increased by 17% which could indicate that using comparable corpora can improve translation for under-resourced languages. The reason why we were interested what amount of text was actually translated was because of the way that the recommendation engine works. It is based on keyword search and named entity recognition, and, if these are not translated, the results might not be good, that is actually relevant to the original text.

Part of the engine is also based on statistic approaches in order to recognise new trending concepts and named entities that can appear in blog posts and news

overnight. The training and learning cycle of a machine translation service has to be short enough to be able to incorporate them into a translation model so that they get properly translated. Because the CC-enhanced method also depends on news crawling, extracting parallel phrases and training translation workers on this data, the learning cycle is longer than ideal (which would be daily integration of new concepts), but it might still be fast enough to be useful when used for Web authoring.

6.6.4 Conclusion

After investigating the importance of translation and specifically MT for Web authoring, we came to the conclusion that translation is much needed and desired on all levels of Web authoring (professional and amateur) and from all perspectives (content creators or content consumers). The quality of MT output is, with some exceptions, still not high enough to be used without human intervention and post-editing, and this is even truer for texts in under-resourced languages and narrow domains. Users in Web authoring use MT output mostly for gisting or as a basis for post-editing.

We described characteristics of Web authoring and user-generated content as well as several requirements that have to be met before successfully applying MT to Web authoring problems.

In our use case, we have shown that MT works well as an intermediate layer between content in under-resourced languages and Web services such as a recommendation engine for related articles which supports only the English language. The recommendation engine returned better suggestions, that is more of the articles were actually related when MT was used to translate texts before feeding them into the recommendation engine.

Even though there are still several obstacles on the path of full utilisation of MT in Web authoring, it already benefits users by helping them bridge the language gap when they are either searching for information, participating in social media networks or enriching their blog posts that are written in an under-resourced language.

6.7 Systems for Computer-Aided Translation

Although the quality of MT systems has been criticised a lot, due to a growing pressure on efficiency and cost reduction, MT receives more and more interest from the localisation industry.

Different aspects of post-editing and machine translatability have been researched since the 1990s (a comprehensive overview has been provided by O'Brien (2005)). Several productivity tests have been performed in translation and localisation industry settings at Microsoft (Schmidtke 2008), Adobe (Flournoy and Duran 2009), Autodesk (Plitt and Masselot 2010), and Tilde (Skadiņš et al. 2011). In all these tests,

authors report productivity increase. However, in many cases, they also indicate significant performance differences in the various translation tasks. Increase of the error score for translated texts is also reported.

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

In this section, we demonstrate 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.
2. If comparable corpora are large enough and can be classified as strongly comparable, then the trained SMT systems applied in the localisation process increase the productivity of human translators.

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 the 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%. This section is based on the publication by Pinnis et al. (2013).

6.7.1 Collecting and Processing a 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 software manuals, etc.). The corpus was acquired in an artificial fashion in order to simulate a strongly comparable narrow domain corpus (i.e. a corpus containing overlapping content in a significant proportion).

To get more data for our experiments, we used two different approaches in the creation of a 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 into chunks of less than 100 paragraphs per document and aligned at document level with the *DictMetric* tool, which is decribed in Chap. 2. As a very large number of alignments were 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 into 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), which is decribed in Chap. 3. The Web corpus consists of 232,665 unique English and 96,573 unique Latvian sentences. The parallel data contained 1,257,142 sentence pairs before pollution.

The statistics of the English–Latvian comparable corpus are given in Table 6.37. Note that the second part of the corpus accounts for 22,498 document pairs.

The parallel sentence extractor *LEXACC*, which is decribed in Chap. 5, 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 tokenised (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 6.38 shows information about data extracted from both parts of the corpus using the selected thresholds.

We applied the ACCURAT Toolkit to acquire in-domain bilingual term pairs from the comparable corpus following the process thoroughly described in Pinnis et al. (2012b). 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 6.39.

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 the 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%).

Table 6.37 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

Table 6.38 Extracted semi-parallel sentence pairs

Corpus part	Threshold	Unique sentence pairs
First part	0.6	9720
Second part	0.35	561,994

Table 6.39 Term tagging and mapping statistics

	Unique monolingual terms		Mapped term pairs	
Corpus part	English	Latvian	Before filtering	After filtering
First part	127,416	271,427	847	689
Second part	415,401	2,566,891	3501	3393

As can be seen in Table 6.39, only a small part of the monolingual terms were mapped. However, this amount of mapped terms was sufficient for SMT system adaptation as described below. 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.

6.7.2 Building SMT Systems

We used the LetsMT! platform (Vasiļjevs et al. 2012) based on the Moses tools (Koehn et al. 2007) 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 the 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) (Bertoldi et al. 2009) using in-domain (IT domain) randomly selected tuning data containing 1837 unique sentence pairs.

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.

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 6.40.

Table 6.40 shows that there was some sentence pair overlap between the DGT-TM corpus 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.

Table 6.40 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	3594	3565

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 the extracted semi-parallel sentences from the in-domain comparable corpus was used to build the in-domain language model.

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 (Pinnis and Skadiņš 2012) 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 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.

6.7.3 Automatic and Comparative 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 6.41.

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.

For the system comparison, we used the same test corpus as for automatic evaluation and compared the baseline system against the adapted system. Figure 6.13

Table 6.41 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 adapted system	No	56.28	9.1805	43.23	0.3998
	Yes	54.81	8.9349	45.04	0.3499
Final adapted system	No	56.66	9.1966	43.08	0.4012
	Yes	55.20	8.9674	44.74	0.3514

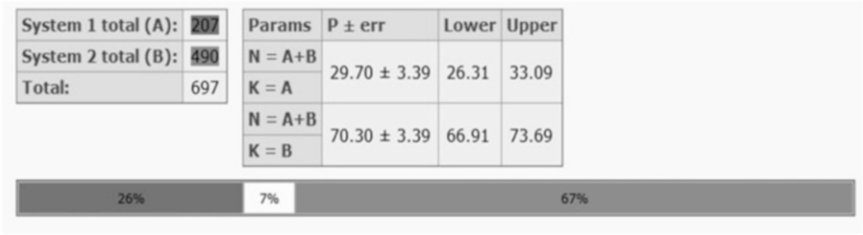

Fig. 6.13 System comparison by total points (System 1—baseline, System 2—adapted system)

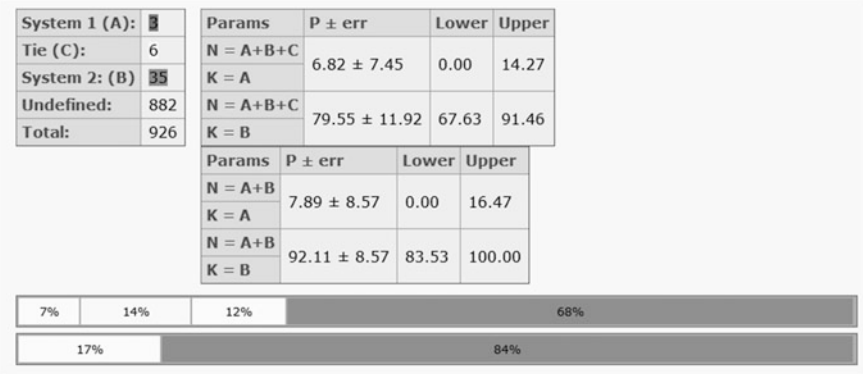

Fig. 6.14 System comparison by count of the best sentences (System 1—baseline, System 2—adapted system)

summarises the human evaluation results using the evaluation method described in Skadiņš et al. (2010). From 697 evaluated sentences, output of the improved SMT system was chosen as a better translation in 490 cases (70.30 ± 3.39%), while users preferred the translation of the baseline system in 207 cases (29.70 ± 3.39%). This allows us to conclude that for IT domain texts, the adapted SMT system provides better translations than the baseline system.

Figure 6.14 illustrates the evaluation on sentence level: we can reliably say that the adapted SMT system provides a better translation for 35 sentences, while users preferred the translation of the baseline system for only 3 sentences. 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.

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

6.7.4.1 Evaluation Set-Up

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 7700 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 the 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. The individual productivity of each translator was measured and compared against his or her own productivity. The average productivity for all of the translators has been calculated using the following formula (6.2):

$$\text{Productivity (scenario)} = \frac{\sum_{\text{Text}=1}^{N} \text{Adjusted words(Text, scenario)}}{\sum_{\text{Text}=1}^{N} \text{Actual time(Text, scenario)}}. \tag{6.2}$$

Usage of MT suggestions in addition to TMs increased the productivity of the translators on average from 503 to 572 words per hour (see Table 6.42). There were significant differences in the results of different translators from a performance increase by 35.4% to decreased performance by 5.9% for one of the translators. Analysis of these differences requires further studies, but they are most likely caused by working patterns and skills of individual translators.

Table 6.42 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

Table 6.43 Quality grades based on error scores

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

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

Editors also calculated an error score for every translation task by counting identified errors and applying a weighted multiplier based on the severity of the error type:

$$\text{ErrorScore} = \frac{1000}{n} \sum_i w_i e_i, \tag{6.3}$$

where n is the number of words in the translated text, e_i is the 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 is assigned a translation quality grade (*Superior*, *Good*, *Mediocre*, *Poor*, or *Very poor*) (Table 6.43).

6.7.4.2 Results

The overall error score (shown in Table 6.44) 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 6.44 Localisation task error score results

Translator	Scenario	Accuracy	Language quality	Style	Terminology	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

6.7.5 Discussion

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 led 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 the experiment should be carried out with more participants for more justified results. It would also be useful to further analyse the correlation between the regular productivity of a translator and the impact on productivity by adding MT support.

Error rate analysis shows that in 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.

References

Abdul-Rauf, S., & Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. *Proceedings of the 12thConference of the European Chapter of the Association for Computational Linguistics* (pp. 16–23), Athens, Greece.

Abdul-Rauf, S., & Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation, 25*(4), 341–375.

Aleksić, V., & Thurmair, Gr. (2011). Personal Translator at WMT 2011. *Proceedings of the WMT Edinburgh*, UK.

Babych, B., & Hartley, A. (2008). Sensitivity of automated MT evaluation metrics on higher quality MT output: BLEU vs task-based evaluation methods. *Proceedings of LREC*, Marrakech.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL 2005)*, June 2005, Michigan.

Bertoldi, N., Haddow, B., & Fouet, J. B. (2009). Improved minimum error rate training in moses. *The Prague Bulletin of Mathematical Linguistics, 91*, 7–16.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., & Monz, C. (2018). *Findings of the 2018 Conference on Machine Translation (WMT18)* (pp. 272–303). WMT (shared task) 2018.

Bontchev, B., & Vassileva, D. (2009). Courseware authoring for adaptive e-learning. *Proceedings of the 2009 International Conference on Education Technology and Computer (ICETC '09)* (pp. 176–180). IEEE Computer Society, Washington, DC.

Bulterman, D. C. A., & Hardman, L. (2005). Structured multimedia authoring. *ACM Transactions on Multimedia Computing, Communication and Applications, 1*, 89–109.

Callison-Burch, Ch., Koehn, Ph., Monz, Ch., & Schroeder, J. (2009). Findings of the 2009 workshop on statistical machine translation. *Proceedings of the 4th Workshop on SMT*, Athens.

Capuano, N., Pierri, A., Colace, F., Gaeta, M., & Mangione, G. R. (2009). A mash-up authoring tool for e-learning based on pedagogical templates. *Proceedings of the First ACM International Workshop on Multimedia Technologies for Distance Learning (MTDL '09)* (pp. 87–94). ACM, New York, NY.

Carrera, J., Beregovaya, O., & Yanishevsky, A. (2009). *Machine Translation for Cross-Language Social Media*. Accessed April 23, 2013 from http://www.promt.com/company/technology/pdf/machine_translation_for_cross_language_social_media.pdf

Clark, E., & Araki, K. (2011). Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. *Procedia – Social and Behavioral Sciences, 27*, 2–11.

Deltour, R., & Roisin, C. (2006). The limsee3 multimedia authoring model. *Proceedings of the 2006 ACM Symposium on Document Engineering (DocEng '06)* (pp. 173–175). ACM, New York, NY.

Désilets, A., Gonzalez, L., Paquet, S., & Stojanovic, M. (2006). Translation the Wiki Way. *The Conference Wiki of the 2006 International Symposium on Wikis*. Odense, Denmark.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research (HLT 2002)* (pp. 138–145), San Diego.

Escudero, H., & Fuentes, R. (2010). Exchanging courses between different Intelligent Tutoring Systems: A generic course generation authoring tool. *Knowledge-Based Systems, 23*(8), 864–874.

Flournoy, R., & Duran, C. (2009). Machine translation and document localization at Adobe: From pilot to production. *Proceedings of the Twelfth Machine Translation Summit*, Ottawa, Canada.

Flournoy, R., & Rueppel, J. (2010). One technology: Many solutions. *AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, CO, 6p.

Forcada, M. (2006). Open-source machine translation: An opportunity for minor languages. *5th SALTMIL Workshop on Minority Languages* (pp. 1–7).

Garcia, I. (2009). Beyond translation memory: Computers and the professional translator. *The Journal of Specialised Translation, 12*, 199–214.

Hamon, O., Popescu-Belis, A., Choukri, K., Dabbadie, M., Hartley, A., Mustafa El Hadi, W., et al. (2006). CESTA: First conclusions of the technolangue mt evaluation campaign. *Proceedings of the LREC*, Genova, Italy.

Hewavitharana, S., & Vogel, S. (2008). Enhancing a statistical machine translation system by using an automatically extracted parallel corpus from comparable sources. *Proceedings of the Workshop on Comparable Corpora, LREC'08* (pp. 7–10).

Hovy, E., King, M., & Popescu-Belis, A. (2002). Principles of context-based machine translation evaluation. *Machine Translation, 17*(1), 43–75.

Hutchins, J. (2003). *Machine translation and computer-based translation tools: What's available and how it's used*. A New Spectrum of Translation Studies. University of Valladolid.

Intel Corporation. (2012). *Enabling Multilingual Collaboration through Machine Translation (IT@Intel White Paper).* Accessed March 30, 2013 from http://www.intel.com/content/www/us/en/it-management/intel-it-best-practices/enabling-multilingual-collaboration-through-machine-translation.html

Irvine, A., & Callison-Burch, Ch. (2013). Combining bilingual and comparable corpora for low resource machine translation. *Proceedings of the Eighth Workshop on Statistical Machine Translation* (pp. 262–270).

Jiang, J., Way, A., & Haque, R. (2012). Translating user-generated content in the social networking space. *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2012)*, San Diego, CA.

King, M., Popescu-Belis, A., & Hovy, E. (2003). FEMTI: Creating and using a framework for MT evaluation. *Proceedings of MT Summit*, New Orleans.

Koehn, P., & Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. *Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*.

Lewis, W., Wendt, C., & Bullock, D. (2010). Achieving domain specificity in SMT without overt siloing. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*.

Lu, B., Jiang, T., Chow, K., & Tsou, B. K. (2010). Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora* (pp. 42–48), Valletta, Malta.

Mehm, F., Reuter, C., Göbel, S., & Steinmetz, R. (2012). Future trends in game authoring tools. *Entertainment Computing-ICEC 2012* (Vol. 7522, pp. 536–541),Springer, Heidelberg.

Mitchell, L., & Roturier, J. (2012). Evaluation of machine-translated user generated content: A pilot study based on user ratings. *Proceedings of the 16th EAMT Conference*, 28–30 May 2012, Trento, Italy.

Mugwanya, R., & Marsden, G. (2010). Mobile learning content authoring tools (MLCATs): A systematic review. *Proceedings E-Infrastructures and E-Services on Developing Countries – Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (pp. 20–31).

Müller, W., Iurgel, I., Otero, N., & Massler, U. (2010). Teaching English as a second language utilizing authoring tools for interactive digital storytelling. *ICIDS'10 Proceedings of the Third Joint Conference on Interactive Digital Storytelling* (pp. 222–227).

Munteanu, D., & Marcu, D. (2006). Improving machine translation performance by ex-ploiting non-parallel corpora. *Computational Linguistics, 31*(4), 477–504.

Najeh, H., Kolovratnik, D., Vaeyrynen, J., Steinberger, R., & Varga,D. (2014). DCEP-digital corpus of the European parliament. *Proceedings of LREC 2014 (Language Resources and Evaluation Conference)* (pp. 3164–3171).

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

Och, F. J. (2003) Minimum error rate training in statistical machine translation. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (Vol. 1, pp. 160–167).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* (pp. 311–318).

Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., & van Genabith, J. (2012). Domain adaptation of statistical machine translation using web-crawled resources: A case study. *Proceedings of the EAMT 2012*, Trento, Italy.

Pinnis, M. (2012). Latvian and lithuanian named entity recognition with TildeNER. *Proceedings of LREC 2012*, 21–27 May, 2012, Istanbul, Turkey.

Pinnis, M., & Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains –What Works and What Not. *Baltic HLT2012*.

Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., et al. (2012a). Toolkit for multi-level alignment and information extraction from comparable corpora. *Proceedings of ACL 2012, System Demonstrations Track*, Jeju Island, Republic of Korea, 8–14 July 2012.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., & Gornostay, T. (2012b). Term extraction, tagging and mapping tools for under-resourced languages. *Proceedings of the 10th Conference on Terminology and Knowledge Engineering*, Madrid, Spain.

Pinnis, M., Skadiņa, I., & Vasiļjevs, A. (2013). Domain adaptation in statistical machine translation using comparable corpora: Case study for english latvian it localisation. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics CICLING 2013*.

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

Popescu-Belis, A. (2008). Reference-based vs. task-based evaluation of human language technology. Proceedings of LREC.

Rirdance, S., & Vasiljevs, A. (Eds.). (2006). *Towards consolidation of European terminology resources. Experience and recommendations from EuroTermBank project*. Riga: EuroTermBank Consortium.

Roturier, J., & Bensadoun, A. (2011). Evaluation of MT systems to translate user generated content. *Proceedings of Machine Translation Summit XIII* (pp. 244–251), Xiamen, China.

Scherp, A., & Boll, S. (2005). Context-driven smart authoring of multimedia content with xSMART. *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)* (pp. 802–803). ACM, New York, NY.

Schmidtke, D. (2008). *Microsoft office localization: Use of language and translation technology*. Available at: http://www.tm-europe.org/files/resources/TM-Europe2008-Dag-Schmidtke-Microsoft.pdf

Schwenk, H., & Koehn, P. (2008). Large and diverse language models for statistical machine translation. *IJCNLP2008*.

Skadiņa, I., Aker, A., Giouli, V., Tufis, D., Gaizauskas, R., Mieriņa, M., et al. (2010). A collection of comparable corpora for under-resourced languages. *Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications*(Vol. 219, pp. 161–168), IOS Press.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlič, M., et al. (2012). Collecting and using comparable corpora for statistical machine translation. *Proceedings of LREC'12* (pp. 438–445), Istanbul, Turkey, 21–27 May 2012.

Skadiņš, R., Goba, K., & Šics, V. (2010). Improving SMT for baltic languages with factored models. *Proceedings of the Fourth International Conference Baltic HLT 2010* (pp. 125–132), October 7–8, 2010, Riga, Latvia.

Skadiņš, R., Puriņš, M., Skadiņa, I., & Vasiļjevs,A. (2011). Evaluation of SMT in localization to under-resourced inflected language. *Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011* (pp. 35–40), May 30–31, 2011, Leuven, Belgium.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. Proceedings of Association for Machine Translation in the Americas.

Snover, M., Madnani, N., Dorr, B., & Schwartz, R. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. *Proceedings of WMT09*.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 11–21.

Ştefănescu, D., Ion, R., & Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)* (pp. 137–144), Trento, Italy.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., TufisD., et al. (2006). The jrcacquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

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, 21–27 May 2012.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., et al. (2014). An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation Journal (LRE), 48*(4), 679–707.

Su, F., & Babych, B. (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-) parallel translation equivalents. *Proceedings of the EACL'12 Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRBMT) and Hybrid Approaches to Machine Translation (HyTra)* (pp. 10–19), Avignon, France, 23–27 April 2012.

Thurmair, Gr., & Aleksić, V. (2012). Creating term and lexicon entries from phrase tables. *Proceedings of the EAMT 2012*,Trento, Italy.

Tiedemann, J. (2009). News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing* (Vol. V, pp. 237–248). Amsterdam/ Philadelphia: John Benjamins.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.

Tyers, F., & Alperen, M. (2010). South-East European Times: A parallel corpus of Balkan languages. *Proceedings of Workshop "Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages"*.

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: A cloud-based platform for do-it-yourself machine translation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL2012)* (pp. 43–48), Jeju, Republic of Korea, 10 July 2012, System Demonstrations.

Watson, C., Li, F. W. B., & Lau, R. W. H. (2010). A pedagogical interface for authoring adaptive e-learning courses. *Proceedings of the Second ACM International Workshop on Multimedia Technologies for Distance Learning (MTDL '10)* (pp. 13–18). ACM, New York, NY.

White, J., O'Connell, T., & O'Mara, F. (1994). The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas* (pp. 193–205). Columbia.

Xu, J., Zens, R., & Ney, H. (2006) Partitioning parallel documents using binary segmentation. *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Statistical Machine Translation* (pp. 78–85), New York City, NY, June 2006.

Xu, J., Deng, Y., Gao, Y., & Ney, H. (2007) Domain dependent machine translation. *Proceedings of the Machine Translation Summit XI*, Copenhagen, Danmark, September 2007.

Zhang, X. (2011). Two-level parallel text extraction from comparable corpora. Diploma thesis of Univeristy of Saarland.