Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)
© 2012 The Authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-133-5-176

MT Adaptation for Under-Resourced Domains – What Works and What Not

Mārcis PINNIS[1] and Raivis SKADIŅŠ[2]
Tilde, Latvia

Abstract. In this paper the authors present various techniques of how to achieve MT domain adaptation with limited in-domain resources. This paper gives a case study of what works and what not if one has to build a domain specific machine translation system. Systems are adapted using in-domain comparable monolingual and bilingual corpora (crawled from the Web) and bilingual terms and named entities. The authors show how to efficiently integrate terms within statistical machine translation systems, thus significantly improving upon the baseline.

Keywords. statistical machine translation, domain adaptation, comparable corpora

Introduction

This paper focuses on a very practical aspect of statistical machine translation (SMT) – tailoring it to a particular narrow domain. The state-of-the-art SMT has reached a level when it can be used by professional translators to improve productivity (for example, see [1]), but to train practically usable domain specific SMT systems we need a significant amount of parallel and monolingual corpora [2][3][1]. In this paper we are searching for methods allowing us to build a domain specific SMT system even if we have a very limited in-domain parallel corpus consisting of just a few thousand parallel sentences.

If we do not have a big domain specific parallel corpus we can look for other resources that could compensate for it. In this paper we show how we can benefit from in-domain texts in the Web, e.g., how we can collect or crawl an in-domain comparable corpus from the Web and how we can use it to build domain specific SMT systems. We are showing how general out-of-domain SMT systems can be tailored using data extracted from the in-domain comparable corpus. Particularly we are dealing with domain specific terminology and named entities (NE). 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

[1] Corresponding author: Tilde, Vienības gatve 75a, Riga, Latvia; E-mail: marcis.pinnis@tilde.lv
[2] Corresponding author: Tilde, Vienības gatve 75a, Riga, Latvia; E-mail: raivis.skadins@tilde.lv

adapted SMT system quality changes are evaluated in respect to a general out-of-domain baseline system. The process is thoroughly described in the further sections.

1. Baseline System

We start our experiments with the creation of an English-Latvian baseline system. In the experiments we assume that the following data is available beforehand:

- a relatively large out-of-domain parallel corpus. For this paper we used the publicly available DGT-TM[3] English-Latvian parallel corpus (release of 2007). The corpus consists of 804,501 unique parallel sentence pairs and 791,144 unique Latvian sentences. The monolingual corpus is used for language modeling.
- 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 is split in two sets - tuning and evaluation. The tuning set and the evaluation set consist of 1,745 and 872 unique sentence pairs from the automotive domain. All systems were tuned with minimum error rate training (MERT [4]) using the in-domain tuning set and evaluated on the evaluation set.

For MT system training (including the baseline system) we use the *LetsMT!* [5] Web-based platform for SMT system creation. The *LetsMT!* platform is built upon the state-of-the-art *Moses* [6] SMT *experiment management system* (*EMS*).

The baseline system's results using different automatic evaluation methods (BLEU [7], NIST [8], TER [9], and METEOR [10]) are given in Table 1.

Table 1 Baseline system's 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

2. Initial Extraction and Alignment of Terms and Named Entities

The first step in our SMT system adaptation technique is acquisition of translated 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 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*) [11] and for named entities – *TildeNER* [12] for Latvian and

[3] The DGT Multilingual Translation Memory of the Acquis Communautaire: DGT-TM (available at: http://langtech.jrc.ec.europa.eu/DGT-TM.html).

OpenNLP[4] 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 allow also 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.

After the symmetric alignment we align also 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. 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 *FMC* [13] crawler. By using bilingually aligned seed terms we ensure that the crawled corpora will be comparable and within one domain for both English and Latvian languages. As the aligned seed terms may contain also 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) [14] scores of words from general (broad) domain corpora (for instance, the whole Wikipedia and current news corpora) to weigh the specificity of a phrase. We rank each bilingual phrase using the following equation:

$$R(p_{src}, p_{trg}) = min\left(\sum_{i=1}^{|p_{src}|} IDF_{src}(p_{src}(i)), \sum_{j=1}^{|p_{trg}|} IDF_{trg}\left(p_{trg}(j)\right)\right) \quad (1)$$

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

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

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.

Additionally 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 2.

Table 2 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	6,155	1,664,403	15,776,967	**271,327**	4,290,213

Both monolingual corpora were aligned in the document level using *DictMetric* [15], a tool that 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 3.

Table 3 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	5,461	247,846	3,939,921

4. 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 tag only terms as the precision of named entity mapping without a phrase table is well below 90% and this would create unnecessary noise in the extracted data for SMT adaptation. Then using the document alignment information of the comparable corpus we map terms bilingually using the *TerminologyAligner* (*TEA*) [15][11] tool with a translation confidence score threshold of 0.7 (with a precision of 90% and higher [11]). In total 369 in-domain term pairs were extracted from the bilingual comparable corpus.

It is possible to use these newly extracted terms in an iterative comparable corpora collection process, thus bootstrapping also the in-domain translated term pair collection.

However, in this paper we limit corpora collection to only one iteration in order to have a proof-of-concept of the whole SMT system adaptation process.

5. SMT System Adaptation

Following domain adaptation methods suggested in earlier research [2][3] we start the SMT adaptation task by adding an in-domain language model built using the Latvian monolingual comparable corpora 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 trained also 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, but the in-domain language model may not contain some general language phrases that the broad domain corpus has (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 adding also 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 6,767 unique in-domain automotive term pairs from EuroTermBank[5]. Then we trained an SMT system (named *Int_LM+ETB_Terms*) 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 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 [16] 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 [15]) we extracted 6,718 and 678 unique sentence pairs with two parallelism confidence score thresholds 0.51 and 0.35 (the thresholds were based on previous evaluation on comparable news domain corpora). These sentence pairs were then added to the available parallel data and the in-domain

[5] EuroTermBank - the largest free online terminology resource (http://www.eurotermbank.com/)

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 manually analysing the MT output of *Int_LM+LEXACC_0.35* in comparison with the baseline system, it is 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 improving SMT quality, but 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. We have shown in [17] that in cases where large strongly comparable in-domain corpora are available, the pseudo-parallel sentences extracted from the corpora (up to 500,000 sentence pairs and more) can achieve a translation quality increase of up to five times in comparison to the baseline system. The challenge, however, is finding such in-domain strongly comparable corpora.

So far in our experiments only the in-domain language model helps distinguishing 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 only 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 on both sides (in both languages) contains a term pair from the term list.

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 hypothesis 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* (using additionally the 369 term pairs from the comparable corpora) to term-aware phrase tables. After tuning with MERT two new systems were created. The system *Int_LM+T_Terms+6th* achieves 13.19 BLEU points and the system *Int_LM+T&CC_Terms+6th* achieves 13.61 BLEU point (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 hypotheses selection for in-domain terminology.

Complete results of the previously described automotive domain systems are shown in Table 4 ("*CS*" stands for "*Case Sensitive*" evaluation).

To show that improvements in SMT quality are consistent also 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 5) show a relative BLEU increase of 8.8% and 14.9% for the system without the 6th feature and with the 6th feature over the baseline. As more data creates higher ambiguity, the 6th feature allows increasing the results significantly more than in the previous experiment. This shows the potential of the method when applied on larger corpora.

Table 4 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 5 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	4.8448	4.6934	73.80	75.12	0.2098	0.1651
Big_Int_LM+T&CC_Terms	17.24	16.12	5.0020	4.8278	72.16	73.59	0.2163	0.1717
Big_Int_LM+T&CC_Terms+6th	**18.21**	**17.08**	**5.1476**	**4.9626**	**70.22**	**71.62**	**0.2191**	**0.1747**

Conclusion

In this paper we have presented techniques for SMT domain adaptation utilizing bilingual terms and bilingual comparable corpora collected from the Web. The experiment results 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 an SMT system quality improvement of 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.

Acknowledgements

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement n° 248347. The research within the project TaaS leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement n° 296312. This work has been supported by the European Social Fund within the project «*Support for Doctoral Studies at University of Latvia*».

References

[1] 01Skadiņš, R., Puriņš, M., Skadiņa, I., and Vasiljevs, A. Evaluation of SMT in localization to under-resourced inflected language. In Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011, 2011, p. 35-40, Leuven, Belgium.

[2] 02Koehn, P. and Schroeder, J. Experiments in domain adaptation for statistical machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, 2007, Prague.

[3] 03Lewis, W., Wendt, C. and Bullock, D., Achieving Domain Specificity in SMT without Overt Siloing. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), 2010.

[4] 04Bertoldi, N., Haddow, B., Fouet, J.B. Improved Minimum Error Rate Training in Moses, The Prague Bulletin of Mathematical Linguistics, Vol. 91 (2009), p. 7-16, Prague, Czech Republic.

[5] 05Vasiljevs, A., Gornostay, T. and Skadins, R. LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. In: Proceedings of the Fourth International Conference Baltic HLT 2010, 2010, p. 133-140, Riga, Latvia.

[6] 06Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, June 2007, Prague, Czech Republic.

[7] 07Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. BLEU: a method for automatic evaluation of machine translation. In Proceedings of ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, 2002, p. 311-318.

[8] 08Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research (HLT 2002), 2002, p. 138-145, San Diego, USA.

[9] 09Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of Association for Machine Translation in the Americas, 2006.

[10] 10Banerjee, S. and Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: 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, USA.

[11] 11Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M. and Gornostay, T. Term Extraction, Tagging and Mapping Tools for Under-Resourced Languages. In: Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), 20-21 June, 2012, Madrid, Spain.

[12] 12Pinnis, M. Latvian and Lithuanian Named Entity Recognition with TildeNER. In Proceedings of LREC 2012, 21-27 May, 2012, Istanbul, Turkey.

[13] 13ACCURAT D3.5. Tools for building comparable corpus from the Web, version 3.0, 29th June, 2012 (http://www.accurat-project.eu/), 46 pages.

[14] 14Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, volume 28, 1972, p. 11-21.

[15] 15ACCURAT D2.6 (2011). Toolkit for multi-level alignment and information extraction from comparable corpora, version 3.0. 29th June, 2012 (http://www.accurat-project.eu/), 164 pages.
[16] 16Ştefănescu, D., Ion, R. and Hunsicker, S. Hybrid Parallel Sentence Mining from Comparable Corpora. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012), 2012, p. 137-144, Trento, Italy.
[17] 17ACCURAT D5.4. Report on requirements, implementation and evaluation of usability in application for software localization, version 1.0. 29th June, 2012 (http://www.accurat-project.eu/), 38 pages.