SMT systems for less-resourced languages based on domain-specific data

Lene Offersgaard[1,] **Dorte Haltrup Hansen**[2]
[1, 2] University of Copenhagen, Center for Language Technology
Njalsgade 140, DK-2300 Copenhagen
E-mail:[1]leneo@hum.ku.dk, [2]dorteh@hum.ku.dk

Abstract

In this paper we show that good SMT systems for less-resourced languages can be obtained by using even small amounts of high quality domain-specific data. We suggest a method to filter newly collected data for parallel corpora, using the internal alignment scores from the aligning process. The filtering process is easy to use and is based on open-source tools. The domain-specific data are used in combination with other public available resources for training SMT systems. Automatic evaluation shows that relatively small amounts of newly collected domain-specific data result in systems with promising BLEU scores in the range of 52.9 to 60.9.
The LetsMT! platform is used to create the presented machine translation systems, where the flexible platform allows uploading the user's own data for training. The paper shows that the platform is a promising way of making SMT systems available for less-resourced languages.

Keywords: SMT, less-resourced languages, domain-specific SMT

1 Introduction

LetsMT! is an EU-project[1] with the aim of building a platform for user tailored machine translation and online sharing of training data. Here we report on resent results of training and evaluating SMT systems for three less-resourced European languages within the LetsMT! platform, all systems based on newly collected domain-specific data. When data are collected from new sources it is a challenge to ensure good parallel data quality. In this paper we suggest a method for filtering the collected data, using alignment scores. The filtered data are then used in combination with other public available resources for training SMT systems. The systems are trained in the LetsMT! platform, which enables the Moses SMT software to do the training with in-domain and out-of-domain language models. Automatic evaluation shows that relatively small amounts of good quality domain-specific data result in systems with promising evaluation scores. In this paper we therefore focus on the process from data collection, data filtering to the SMT system training and evaluation within the LetsMT! platform, a platform which gives the opportunity for new users to create their own domain specific SMT system of fairly good quality by means of limited quantity of in-domain data.

2 LetsMT! platform

The LetsMT! platform[2] allows users to upload their own data into a repository, which converts, store and handle data in a safe and functional way to prepare data for training standard SMT engines (Tiedemann et al. 2012). From an easy-to-use web-interface registered users can configure an SMT engine based on a combination of large public resources and other resources uploaded to the platform - either by the user itself or other users. An efficient cloud-based training can then be carried out based on the Moses SMT software[3] with the in-domain and out-of domain data handling described in Koehn and Schroeder, 2007 The LetsMT! platform also allows for integration in SDL Trados - integration with other CAT tools is under development, enabling easy use of the LetsMT! system for localization. For testing purpose and minor translation tasks a web-interface is available.

3 Domain issues in SMT

In LetsMT! data has been collected for a number of subject domains. Our assumption is the quality of automatic translation increases if the systems are trained on domain-specific data. In (Pecina et al., 2011) an approach of tuning existing general-domain systems with domain-specific data did not seem promising. In (Offersgaard et al., 2008) systems were trained on domain-specific data, but here a method weighing a domain-specific phrase table higher than a more general phase table showed an increase in BLEU and TER scores. In this paper we focus on the options given in the LetsMT! platform (Koehn and Schroeder, 2007), where in-domain and out-of-domain language models are weighted.

An important issue is to classify the data in named subject domains. In an ideal world it would be preferable if collected data could be classified in the same large-scale general subject classification system. Not only would it ease the identification of consistent and representative bilingual training data, it would also, via the fine-grained subject classification, increase the probability that the lexical coverage of a given SMT-system would be tuned for the texts to be translated. But unfortunately a large

[1] LetsMT! is supported by the European Commission's ICT Policy Support Programme and is running from Mar. 1st 2010 until Aug. 31 2012

[2] See http://letsmt.eu for theLetsMT! platform

[3] http://www.statmt.org/moses/

universal classification system involves too much administrative work (Rirdance&Vasiljevs, 2006) being a difficult task to classify collected data. In addition, subject classification systems do not take into account possible divergences in the data within the same subject domain, e.g. different companies may have chosen to have different specific company terminologies.

Besides, texts from the same subject domain will make use of very different writing styles in terms of sentence types and varieties in language usage according to the genre of the text. Marketing texts, for instance, may praise the features of the product while manuals focus on strict instructions on how to use the product.
Consequently, in principle it would be preferable to train SMT systems on texts with almost identical writing styles and within the same subject domain.

In LetsMT! we decided to have the limited number of 15 subject domains available. These subject domains include the 10 domains used in TAUS[4].When only a few broad domains are available while uploading data the user can easily select the most appropriate subject domain.

As a supplement to the subject domain specification, the user can also specify text type, a description of the corpus and other metadata for the corpus. This allows users to give detailed information, and to use this information when selecting data for training a specific SMT system.

4 Data collection

The LetsMT! platform gives the opportunity to train domain-specific systems based on data uploaded to the LetsMT! resource repository. The available data in the repository consist of the large and well-known publicly available corpora e.g. Europarl, DGT-TM Acquis Communitare and the Opus corpora, all resources often used for SMT systems. In the LetsMT! platform these resources serve as backbone for training the phrase table and building the language model. In addition to the public available resources domain-specific data for under-resourced languages is collected by the project.

One of these domains is *Business and financial news*. This domain is chosen as a use case for an on-line translation service of financial news into less-resourced languages. The data collected for the domain is annual reports, which have been harvested automatically from a selected list of web sites. Annual reports are mostly freely available on companies' web sites in pdf format.

Another subject domain in focus is *Education* for which administrative documents from Danish Universities were collected, mainly curricula. This use case takes advantage of the LetsMT! plug-in to SDL Trados. Danish universities have an increasing demand for translation of curricula since a large number of courses are now taught in English allowing foreign students an easy access to education in Denmark.

The data collection was not done by web crawling systems but by systematic conduction of relevant web sites to secure high quality of in domain parallel resources.

5 Filtering data

When data is collected automatically noise arises from different sources: the files might be broken or have different content than expected, the translations might not be totally parallel, the layout might have destroyed the text in the pdf-to-xml conversion etc. These factors consequently lead to bad sentence-alignment. Normally large amounts of data ensure to blur bad alignment, but in our set up where only little domain specific data is available, high quality data is required.

As filter we used the alignment types and alignment scores from the HunAligner[5] (Varga et al. 2005). The HunAligner first does a Gale&Church sentence-length based alignment and then builds an automatic dictionary based on this alignment and realigns the text. The aligner produces 0-alignments, when segments have no corresponding segments in the other language, and n:m-alignments, specifying that n segments in the source language correspond to m segments in the target language.

After collection the data were first converted into text, tokenized, converted into xml and aligned by the Uplug tools (Tiedemann, 2002). Then 0-alignments were removed and the average scores calculated for each document. In the filtering process our aims were twofold: we wanted to provide good quality data to the LetsMT! platform and we wanted to find methods for filtering the data automatically.

From the average alignment scores we have done manual inspection of documents with a low average score (< 2). It seemed, however, that this wasn't a sufficient clue for alignment quality. In Pecina et al., 2011 an absolute score of 0.4 was used to filter out bad alignment. Our observations are, however, that especially positive low scores are not reliable while negative scores, high positive scores and average scores for the entire document are more useable. We therefore investigated documents with a high per cent of negative alignments (> 10%). In this case all parallel documents were of a bad quality. We also inspected documents without negative alignments. Absence of negative alignments can either indicate a perfectly parallel translation, an English-English ”translation” (the same file) or empty files. Finally we searched for the English word *the* in the non-English documents to spot false translations or

[4] http://www.translationautomation.com/

[5] http://mokk.bme.hu/resources/hunalign/

language pairs being swopped.

Annual reports	Swedish	Danish	Dutch
Av. score, all reports	2.92	3.1	3.57
Av. score < 2	17 %	8 %	14 %
Neg. scores > 10%	13.5 %	7.7 %	7.4 %
Neg. scores = 0%	4 %	3.5 %	14.8 %
% of documents with mixed languages	1.6 %	4.2 %	2.8 %
% of documents filtered out	16.1 %	14.7 %	19 %

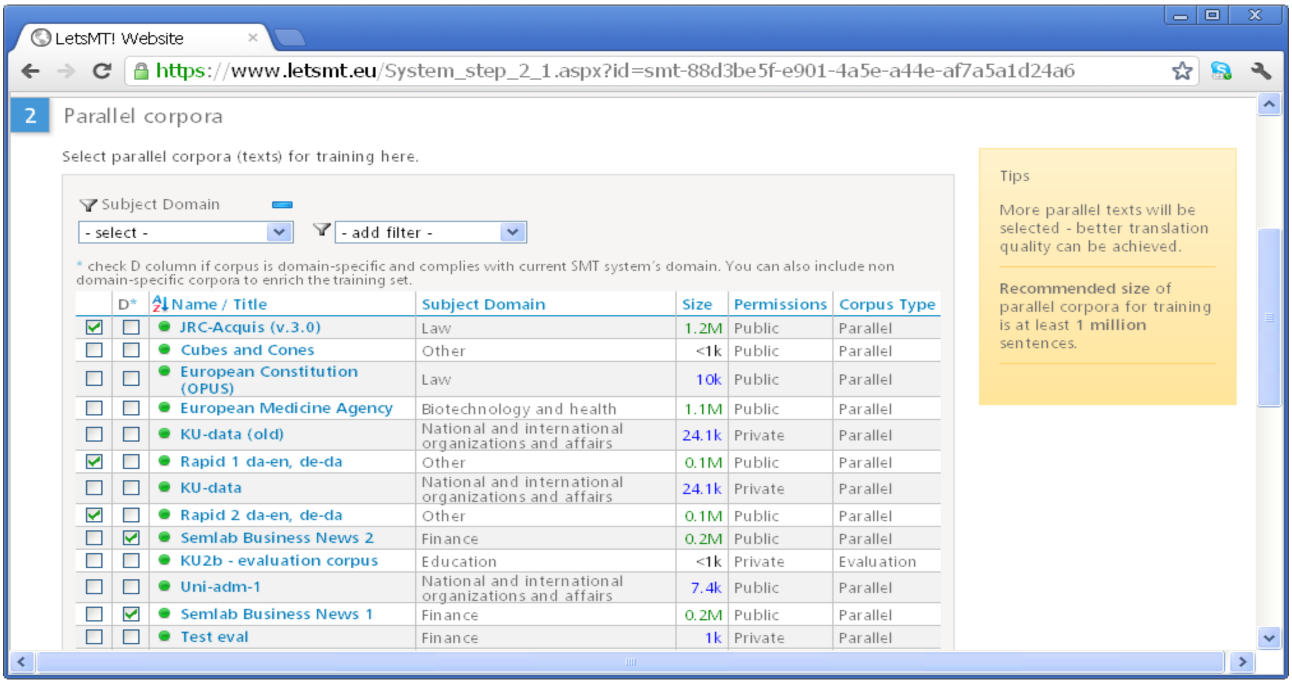
Table 1: Parameters for data filtering.

Table 1 shows the distribution of average alignment scores for Swedish, Danish and Dutch annual reports and the percentage of documents filtered out on the background of the findings.

We suggest that high quality data in terms of being parallel and in-domain, can be filtered by using the negative alignment scores from the HunAligner. Our findings are that positive alignment scores are less reliable than negative scores and that the average percentage of negative scores is a very good indicator for the alignment quality of the document and therefore of the data quality. It is difficult to set a fixed cut-off limit but our manual investigations showed that a threshold of around 10 % negative alignments per document was the upper limit. The table below shows the sizes of the domain-specific corpora after filtering. These corpora are used for training the SMT systems described in the next sections.

Language pair and domain	**Words (English)**
English-Danish Annual reports	3 022 233
English-Dutch Annual reports	5 753 369
English-Swedish Annual reports	11 503 078
Danish-English Education	635 685

Table 2: Size of domain-specific corpora after filtering



LetsMT! Website
https://www.letsmt.eu/System_step_2_1.aspx?id=smt-88d3be5f-e901-4a5e-a44e-af7a5a1d24a6
2

Parallel corpora

Select parallel corpora (texts) for training here.

Subject Domain

- select -

- add filter -

* check D column if corpus is domain-specific and complies with current SMT system's domain. You can also include non domain-specific corpora to enrich the training set.

	D*	Name / Title	Subject Domain	Size	Permissions	Corpus Type
☑	□	● JRC-Acquis (v.3.0)	Law	1.2M	Public	Parallel
□	□	● Cubes and Cones	Other	<1k	Public	Parallel
□	□	● European Constitution (OPUS)	Law	10k	Public	Parallel
□	□	● European Medicine Agency	Biotechnology and health	1.1M	Public	Parallel
□	□	● KU-data (old)	National and international organizations and affairs	24.1k	Private	Parallel
☑	□	● Rapid 1 da-en, de-da	Other	0.1M	Public	Parallel
□	□	● KU-data	National and international organizations and affairs	24.1k	Private	Parallel
☑	□	● Rapid 2 da-en, de-da	Other	0.1M	Public	Parallel
□	☑	● Semlab Business News 2	Finance	0.2M	Public	Parallel
□	□	● KU2b - evaluation corpus	Education	<1k	Private	Evaluation
□	□	● Uni-adm-1	National and international organizations and affairs	7.4k	Public	Parallel
□	☑	● Semlab Business News 1	Finance	0.2M	Public	Parallel
□	□	● Test eval	Finance	1k	Private	Parallel

Tips

More parallel texts will be selected - better translation quality can be achieved.

Recommended size of parallel corpora for training is at least 1 million sentences.

Figure 1: Selecting parallel corpora at the LetsMT! platform

Figure 1 shows how the LetsMT! platform allows the user to select parallel domain-specific and parallel general corpora when training an English-Danish finance SMT system.

6 LetsMT! system training

As reported in section 5 the collected in-domain data are of a relatively small size compared to often suggested amounts of training data for SMT systems. A minimum of 1 M parallel segments and 5 M mono-lingual segments for the language model are normally recommended by LetsMT!.

The LetsMT! Platform enables two ways of applying evaluation and tuning sets to the training process. Either the user can define the sets when configuring the training process or the system can automatically extract sets of 1000 segments from the in-domain training corpora. In both cases - user-defined or automatic - the training data is afterwards cleaned-up for potential overlap between the training data and the selected tuning and evaluation sets. The evaluation sets used for the systems in section 7 and 9 are extracted automatically from the in-domain training data. In section 8 the same evaluation set is used for all systems.

7 Financial SMT systems

As a starting point we have trained three comparable financial SMT systems covering three different language pairs for the financial sub-domain 'Annual reports'. These three systems are trained using both in-domain and out-of-domain data. In table 3 and 4 the amounts of training data are shown. The in-domain data used are the corpora described in section 5. The out-of-domain parallel data for the English-Danish system is a corpus of EU press releases from Rapid, which can be seen as text from a general domain. For the English-Dutch and the English-Swedish systems the EU DGT Acquis corpus was used as out-of-domain data since we did not have a general corpus of original written text for these languages. The monolingual training data are a combination of the target language of the parallel data and the EU DGT Acquis corpus.

System	**In-domain parallel data**	**Out-of-domain parallel data**	**Total**
English-Danish Annual reports I	113 509	194 239	307 748
English-Dutch Annual reports I	307 807	360 449	668 256
English-Swedish Annual reports I	504 572	398 063	902 632

Table 3: Parallel training data (segments)

System	**In-domain mono-lingual data**	**Out-of-domain mono-lingual data**	**Total**
English-Danish Annual reports I	113 509	1 170 532	1 284 041
English-Dutch Annual reports I	307 807	379 225	687 032
English-Swedish Annual reports I	504 572	403 570	908 142

Table 4: Monolingual training data (segments)

The three systems are evaluated using the automatic measures: BLEU (Papineni et al. 2002), Meteor (Denkowski & Lavie 2011) and TER (Snover et al. 2006) (see table 5). The evaluation sets are extracted automatically.

System	**BLEU**	**Meteor**	**TER**
English-Danish Annual reports I	59.75	0.493	48.6
English-Dutch Annual reports I	52.89	0.368	52.3
English-Swedish Annual reports I	55.25	0.384	47.6

Table 5: Evaluation scores for the financial systems

Both the BLEU and the Meteor scores are calculated case-insensitive, to leave out casing issues from the evaluation. Meteor is used with the language independent option, not bringing all Meteor modules into play. Please mark that the TER score indicates the number of edits needed to adjust the translation output according to the reference translation. Therefore a lower TER score is better.

The scores show that all three systems have relative high BLEU and Meteor scores. The TER score reveals that even with these high BLEU scores the number of edits needed to adjust the translation outputs according to the reference translations are substantial – 48% to 52% changes. The evaluation scores cannot be compared among the three systems - and therefore we cannot state that one of the systems is better than the other two systems - as the evaluation corpora are different for the three systems. But we can see that the evaluation scores are very promising for these systems covering the financial sub-domain of 'Annual reports', even with different amount of in-domain and out-of-domain data.

8 More data or in-domain data?

The generally good results for the trained financial systems led us to train additional systems to see which factors made the biggest impact on the translation quality: the amount of data, the domain-specific data or the filtering of data.

For this experiment we focused on the English-Danish annual reports and trained 4 different systems: a baseline system containing only out-of-domain data (the EU DGT Acquis corpus and the EU press releases from Rapid), Annual reports I (as described in section 7), Annual reports II (only in-domain data) and Annual reports III (only in-domain data filtered for bad aligned files).

Systems (En-Da)	**In-domain parallel data**	**Out-of-domain parallel data**	**Total**
Baseline	-	897 548	897 548
Annual reports I	113 509	194 239	307 748
Annual reports II	113 509	-	113 509
Annual reports III	109 644	-	109 644

Table 6: Parallel training data (segments) for the En-Da annuals report systems

Systems (En-Da)	**In-domain mono-lingual data**	**Out-of-domain mono-lingual data**	**Total**
Baseline	-	1 170 532	1 170 532
Annual reports I	113 509	1 170 532	1 284 041
Annual reports II	113 509	-	113 509
Annual reports III	109 644	-	109 644

Table 7: Monolingual training data (segments) for the En-Da annuals report systems

The systems were tested on the same 1000 in-domain segments.

Systems (En-Da)	**BLEU**	**Meteor**	**TER**
Baseline	17.12	0.210	86.2
Annual reports I	59.75	0.493	48.6
Annual reports II	60.04	0.409	49.3
Annual reports III	60.91	0.413	46.8

Table 8: Evaluation scores for the En-Da annuals report systems

The evaluation scores show very clearly that domain-specific data increases the translations quality significantly. It is more surprising that the quality remains at the same level even when only a little amount of in-domain data is used. We believe that this might have to do with the special text type we are dealing with, namely annual reports. The vocabulary and the syntactic structures for this text type are relatively narrow. Finally, the evaluation scores show that filtering out the bad aligned documents gives a small additional improvement in both BLEU and TER even though only 3865 segments were removed.

9 Educational domain

To test if the same kind of quality can be achieved for other subject domains, we trained a system on the relatively small amount of curricula from Danish universities. The results show that with an in-domain parallel corpus containing only 0.6 M words (19,415 segments) and a general parallel corpus containing 526,302 segments, a BLEU score of 56.3 can be achieved.

System	**BLEU**	**Meteor**	**TER**
Danish-English Education with Acquis DGT	56.31	0.408	53.9

Table 9: Evaluation scores for Educational domain

Translators from the translation department on University of Copenhagen have inspected the translated evaluation set and find that the translations are very usable. They are currently evaluating the Danish-English Education system integrated in SDL Trados by the LetsMT! plug-in and report it being a very efficient way to include SMT in their translation workflow.

10 Conclusions

In this paper we describe the process from collection of new domain-specific data for less-resourced languages, filtering the data based on alignment scores, to training systems using the LetsMT! platform. Three systems for the same text type (annual reports) but for three different language pairs (Danish, Swedish, Dutch) were trained. The combination of in-domain and out-of-domain data shows promising automatic evaluation scores with BLEU scores from 52.9 to 55.3. The TER scores are 48% to 52%, revealing that even with the high BLEU scores the output still need quite some editing to match the translation references.

When collecting data from the web, some documents turn out to be lesser parallel as they might look at the first glace. We therefore present a usable method for filtering the collected data based on the negative alignment scores from the HunAligner. Our findings are that the average percentage of negative scores is a very good indicator for the alignment quality of the document. We suggest a cut-off limit of 10% negative alignments per document.

We also investigated the effect of both in-domain data and the amount of data on the translation quality. Results from a baseline system without in-domain data and a system with a combination of all available in-domain data and the same out-of-domain data as the baseline are presented. The system including in-domain data was significantly better than the baseline system with BLEU scores going from 17.1 up to 59.8. Furthermore systems based only on in-domain data – filtered and unfiltered - were trained. Surprisingly the BLEU score remained at the same level for the system with only in-domain data, namely 60.0 compared to 59.8 for the system with the much bigger amount of out-of-domain data. The filtered system showed a small additional improvement with a BLEU of 60.9.

We will conclude by saying that using the LetsMT! platform is a promising way of making SMT systems available for less-resourced languages. Users can now easily create tailored machine translation system taking advantage of the flexible way of including their own data for training SMT systems.

11 Acknowledgements

We want to thank our project partners in the LetsMT! project for collaboration. Especially we would like to thank Thomas Dohmen, SemLab, The Netherlands for web crawling the annual reports and Jörg Tiedeman, University of Uppsala for providing the Uplug tools.

12 References

Anil Kumar Singh and Samar Husain. Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs. In *Proceedings of ACL 2005 Workshop on Parallel Text*. Ann Arbor, Michigan. June 2005

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590-596.

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

Gale, W.A. and K.W. Church. 1994. .A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):pp75-102.

Koehn, P. and Schroeder J. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, 2007,* pages 224–227, Prague, Czech Republic

Lavie, A and Denkowski, M. The METEOR Metric for Automatic Evaluation of Machine Translation, *Machine Translation*, 2010

NIST 2005. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Retrieved 2010-04-17. *Machine Translation Evaluation Official Results.*

Offersgaard, L., Povlsen, C., Almsteen, L., Maegaard, B., Domain specific MT in use. In *Proceedings of the 12th EAMT conference*, 22-23 September 2008, Hamburg, Germany

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

Pecina,P et.a. Towards UsingWeb-Crawled Data for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the15th EAMT conference*, 2011.

Public project report LetsMT! D1.1 *Report on requirements analysis*, 2010, http://project.letsmt.eu

Rirdance, S. Vasiljevs, A,: Towards Consolidation of European Terminilogy Resources. *Experiments and Recommoondations from EuroTermBank Project*. Riga 2006

Snover, M., Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, A Study of Translation Edit Rate with Targeted Human Annotation, *Proceedings of Association for Machine Translation in the Americas*, 2006.

Tiedemann, J. Uplug - a modular corpus tool for parallel corpora. In *Parallel Corpora, Parallel Worlds*, pages 181-197, Rodopi, 2002.

Tiedemann, J, Hansen, D.H., Offersgaard, L., Olsen, S., Zumpe, M. A Distributed Resource Repository for Cloud-Based Machine Translation. . In *Proceedings of LREC 2012*