Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation

Pavel Pecina,
Antonio Toral, Andy Way
School of Computing
Dublin City Universiy
Dublin 9, Ireland
{ppecina,atoral,away}@computing.dcu.ie

Vassilis Papavassiliou,
Prokopis Prokopidis, Maria Giagkou
Institute for Language & Speech Processing
Artemidos 6 & Epidavrou
151 25 Maroussi, Greece
{vpapa,prokopis,mgiagkou}@ilsp.gr

Abstract

This paper reports on the ongoing work focused on domain adaptation of statistical machine translation using domain-specific data obtained by domain-focused web crawling. We present a strategy for crawling monolingual and parallel data and their exploitation for testing, language modelling, and system tuning in a phrase--based machine translation framework. The proposed approach is evaluated on the domains of Natural Environment and Labour Legislation and two language pairs: English–French and English–Greek.

1 Introduction

Performance of a statistical machine translation (SMT) system usually drops when it is applied on data of a different nature than that the system was trained on (Banerjee et al., 2010). As any other machine-learning application, SMT is not guaranteed to perform optimally if the data for training and testing are not identically (and independently) distributed, which is often the case in practice. The main problem is usually vocabulary coverage: specific domain texts typically contain a lot of special vocabulary that is not likely to be found in texts from other domains. Problems are also caused by divergence in style or genre, where the difference is not only in terminology but also in grammar.

In order to achieve optimal performance, an SMT system must be trained on data from the same domain, of the same genre, and the same style as that it is applied on. For many domains, such training resources (monolingual and parallel data) are not available in large enough amounts to train a system of a sufficient quality. However, even small amounts of such data can be used to adapt an existing (general-domain) system to the particular domain (Koehn and Schroeder, 2007). If the data is not available at all, a possible solution is to exploit publicly available data from the web.

In this work, we present a strategy for crawling domain-specific texts from the web and their exploitation for domain-adaptation in a phrase-based statistical machine translation (PB-SMT) framework. At the current stage, we focus on two resources: in-domain parallel data for parameter tuning and in-domain monolingual data for language model training. As part of our approach, we also create domain-specific test sets. The evaluation is carried out on the domains of Natural Environment (*env*) and Labour Legislation (*lab*) and two language pairs: English–French and English–Greek.

The remaining part of the paper is organized as follows. After an overview of related work, we discuss the possibility of adapting a general-domain SMT system to a specific domain by using various types of in-domain data. Then, we describe our baseline SMT system and the web-crawling strategy for monolingual and parallel data. Finally, we report on the results, make conclusions and outline the future directions of our work.

2 Domain adaptation in SMT

Domain adaptation is an active topic in SMT. It was first introduced by Langlais (2002) who integrated in-domain lexicons into the translation model. His work was followed by many others. Eck et al. (2004) presented a language model adaptation technique applying an information retrieval approach based on selecting similar sentences from available training data. Hildebrand et al. (2005) applied the same approach on the translation model. Wu and Wang (2004) and Wu et al. (2005) proposed an alignment adaptation approach to improve domain-specific word alignment. Munteanu and Marcu (2005) automatically extracted in-domain bilingual sentence pairs

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 297–304
Leuven, Belgium, May 2011

languages (L1–L2)	sentence pairs	L1 tokens / vocabulary		L2 tokens / vocabulary	
English–French	1,725,096	47,956,886	73,645	53,262,628	103,436
English–Greek	964,242	27,446,726	61,497	27,537,853	173,435

Table 1: Europarl corpus statistics for relevant language pairs.

from large comparable (non-parallel) corpora to enlarge the in-domain bilingual corpus. Koehn and Schroeder (2007) integrated in-domain and out-of--domain language models as log-linear features in the Moses (Koehn et al., 2007) PB-SMT system with multiple decoding paths for combining multiple domain translation tables. Nakov (2008) combined in-domain translation and reordering models with out-of-domain models also into Moses. In this work, log-linear features were derived to distinguish between phrases of multiple domains by applying data source indicator features. Finch and Sumita (2008) employed a probabilistic mixture model combining two models for questions and declarative sentences with a general model. They used a probabilistic classifier to determine a vector of probability representing class membership.

Domain adaptation of SMT can be approached in various ways depending on the availability of domain-specific data and their type. If the data is available, it can be directly used to improve components of the MT system: word alignment and phrase extraction (Wu and Wang, 2004), language models (Koehn and Schroeder, 2007), and translation models (Nakov, 2008), usually by merging the data with general-domain data or by training new models and using them together with the general--domain ones in the log-linear framework. If the data is not available, it can be extracted from a pool of texts from different domains (Eck et al., 2004; Hildebrand et al., 2005) or even from the web, which is the case in this work. We crawl monolingual data for language models and parallel data to improve parameter tuning. The crawled parallel data is also used to create domain-specific test sets.

3 Baseline system

Our baseline system is MaTrEx, a combination-based multi-engine architecture developed at Dublin City University (Penkale et al., 2010) exploiting aspects of both the Example-based Machine Translation (EBMT) and SMT paradigms. The architecture includes various individual systems: phrase-based, example-based, hierarchical phrase-based, and tree-based MT. In this work, we only exploit the SMT phrase-based component of the system which is based on Moses, a well--known open-source toolkit for SMT. In addition to Moses, MaTrEx provides a set of tools for easy--to-use preprocessing, training, tuning, decoding, postprocessing, and evaluation.

3.1 General-domain data

As for other data-driven MT systems, MaTrEx requires certain data to be trained on, namely parallel data for translation models, monolingual data for language models, and parallel development data for tuning of system parameters. Parameter tuning is not strictly required but has a big influence on system performance. For the baseline system we decided to exploit the widely used data provided by the organizers of the series of SMT workshops (WPT 2005, WMT 2006–2010)[1]: the Europarl parallel corpus (Koehn, 2005) version 5 as training data for translation models and language models, and WPT 2005 test set as the development data for parameter optimization.

The Europarl parallel corpus is extracted from the proceedings of the European Parliament. For practical reasons we consider this corpus to contain general-domain texts. Version 5 released in Spring 2010 includes texts in 11 European languages including all languages of our interest (English, French, and Greek; see Table 1). Note that the amount of parallel data for English and Greek is only about one half of what is available for English and French. Furthermore, Greek morphology is more complex than French morphology so the Greek vocabulary size (count of unique lowercased alphabetical tokens) is much larger than the French one (see Table 1).

The WPT 2005 dev set is a set of 2,000 sentence pairs available in the same languages as Europarl provided by the WPT 2005 workshop organizers as a development set for the translation shared task. Later WMT test sets do not include Greek data.

3.2 System setting

For training the baseline MT system, all training data is tokenized and lowercased using the standard Europarl tools.[2] The original (non-

[1] http://www.statmt.org/

[2] http://www.statmt.org/europarl/

language	dom	websites	docs	sentences	tokens	vocabulary	new vocab.	sample size / accuracy %	
English	*env*	146	505	53,529	1,386,835	33,400	10,276	224	92.9
	lab	150	461	43,599	1,223,697	25,183	6,674	215	91.6
French	*env*	106	543	31,956	1,196,456	36,097	9,485	232	95.7
	lab	64	839	35,343	1,217,945	23,456	5,756	268	98.1
Greek	*env*	112	524	37,957	1,158,980	55,360	17,986	227	97.4
	lab	117	481	34,610	1,102,354	52,887	16,850	219	88.1

Table 2: Web-crawled monolingual data statistics.

lowercased) versions of the target sides of the parallel data are kept for training the Moses recaser. The lowercased versions of the target sides are used for training an interpolated 5-gram language model with Kneser-Ney discounting using the SRILM toolkit (Stolcke, 2002). Translation models are trained on the relevant parts of the Europarl corpus, lowercased and filtered on sentence level; we kept all sentence pairs having less than 100 words on each side and with length ratio within the interval $\langle 0.11,9.0 \rangle$. Minimum error rate training (Och, 2003, MERT) is employed to optimize the model parameters on the development set.

For decoding, test sentences are tokenized, lowercased, and translated by the trained system. Letter casing is then reconstructed by the recaser and extra blank spaces in the tokenized text are removed in order to produce human-readable text.

4 Acquisition of in-domain resources

4.1 Web crawling of monolingual data

Our workflow for acquiring in-domain monolingual data consists of the following steps: focused web crawling, text normalization, language identification, document clean-up and near-duplicate detection. For focused web crawling, we adapted the open-source Combine[3] crawler, which interacts with a text-to-topic classifier. Each web page visited by the crawler is classified as relevant to the domain with respect to a topic definition provided by the user. We used lists of triplets ⟨*term, relevance weight, topic class*⟩ as the basic entities of the topic definition. During crawling, a relevance score s for each web page is calculated as in the formula below (N is the amount of terms in a topic definition; w_i^t is the weight of term i; w_j^l is the weight of location j; n_{ij} is the number of occurrences of term i at location j; l_j is the number of words at location j).

$$s = \sum_{i=1}^{N} \sum_{j=1}^{4} \frac{n_{ij} w_i^t w_j^l}{l_j}$$

[3]http://combine.it.lth.se/

We adopted Ardö's (2005) approach in considering four discrete locations in a web page (*title*, *metadata*, *keywords*, and *plain text*) and experimentally setting the corresponding weights for these locations to 10, 4, 2, and 1. If the score is greater than a predefined threshold, the web page is classified as relevant to the specific domain and the links of the page are extracted. Finally, the crawler follows the extracted links to visit new pages.

To construct the list of triplets of the topic definition, we selected English, French, and Greek terms (both single and multi-word entries) from the domains with identifiers 52 (Natural Environment) and 44 (Employment and Working Conditions) of the Eurovoc thesaurus v4.3.[4] For each language, we extracted 209 terms for the *env* domain and 86 for the *lab* domain. The weights assigned to the terms were signed integers indicating the relevance of each term to a topic-class. Topic-classes correspond to possible sub-categories of the domain.

The other input for the crawler is a list of seed URLs relevant to the domain. The seeds for the *env* domain were selected from relevant lists in the Open Directory Project,[5] a repository maintained by volunteer editors. For the *lab* domain, similar lists were not so easy to find. We therefore adopted a different method, namely using the BootCat toolkit (Baroni and Bernardini, 2004) to create random tuples (i.e. n-combinations of terms) from the terms included in the topic definition. We then ran a query for each tuple on the Yahoo! search engine,[6] kept the first five URLs returned for each query and finally constructed the seed list with these URLs.

Normalization, the next step in the workflow, concerned encoding identification based on the *content_charset* header of each document, and, if needed, conversion to UTF-8. Language identification was performed by a modified version of the n-gram-based Lingua::Identify[7] tool, which was

[4]http://eurovoc.europa.eu/

[5]http://www.dmoz.org/Science/Environment/

[6]http://www.yahoo.com/

[7]http://search.cpan.org/ ambs/Lingua-Identify-0.29/

languages (L1–L2)	dom	websites	documents	sentences all / filtered / sampled / corrected			
English–French	*env*	6	559	16,487	13,840	3,600	3,392
	lab	4	900	33,326	23,861	3,600	3,411
English–Greek	*env*	6	151	4,543	3,735	3,600	3,000
	lab	4	125	3,094	2,707	2,700	2,506

Table 3: Web-crawled parallel data statistics.

used to discard documents not in the targeted language. Web pages often need to be cleaned from "noise" such as navigation links, advertisements, disclaimers, etc. (a.k.a. boilerplate), which are of limited or no use for the purposes of training an MT system. Such noise was removed by the Boilerpipe tool (Kohlschütter et al., 2010).[8] The following step in the workflow involved applying the SpotSigs algorithm (Theobald et al., 2008) to detect and remove near duplicate documents.

The collections consisted of documents originating from as many different web sites as possible, in order to avoid bias to the language of specific sites. Documents originating from bilingual web sites were excluded, as these sites were used for the acquisition of the parallel data (Section 4.2). The only postprocessing steps performed on the monolingual data prior to SMT training were tokenization and sentence boundary identification by the Europarl tools.

The statistics of the data are provided in Table 2. The vocabulary column contains the amount of unique lowercased alphabetical tokens (words) in each data set and the new vocabulary column then shows counts of such tokens not appearing in the Europarl corpus. The ratio of new vocabulary is around 30% for all these data sets, which is encouraging, as by using them a better coverage of in-domain test sets can be expected. To evaluate the crawler's accuracy, we asked two native speakers for each language to classify a sample of the crawled data (selected to achieve at least a $\pm 5\%$ confidence interval at a 95% confidence level) as out-domain or in-domain. The accuracy measured on documents judged as in-domain by both evaluators ranges from 88% to 98% (details in Table 2).

4.2 Web crawling of parallel data

The workflow for acquiring in-domain parallel data consisted of the following steps: first, web sites containing texts in targeted domains and pairs of languages were manually identified from the pool of web sites collected during the phase of monolingual data acquisition (Section 4.1). Pages from those sites were then used as seed URLs and the crawler was constrained to follow only links internal to each site. This constraint was applied in order to force the crawler to stayon the selected multilingual web sites. Each web page visited by the crawler was classified as relevant with respect to a bilingual topic definition. After following the normalization and language identification steps described in Section 4.1, we end up with in-domain EN–FR or EN–EL subsets of websites mirrored locally. The next step concerned using Bitextor (Esplà-Gomis and Forcada, 2010),[9] an open source tool that uses shallow textual features to decide which documents could be considered translations of each other, and to identify pairs of paragraphs from which parallel sentences could be extracted.

The next steps of the procedure aimed at identification of sentence pairs which are likely to be mutual translations. In each paragraph pair we applied the following steps: identification of sentence boundaries by the Europarl sentence splitter, tokenization by the Europarl tokenizer, and sentence alignment by Hunalign,[10] a widely used tool for automatic identification of parallel sentences in parallel texts. For each sentence pair identified as parallel, Hunalign provides a score which reflects the level of parallelness, the degree to which the sentences are mutual translations. We manually investigated a sample of sentence pairs extracted by Hunalign from the pool data for each domain and language pair (45–49 sentence pairs for each language pair and domain), by relying on the judgement of native speakers, and estimated that sentence pairs with a score above 0.4 are of a good translation quality. In the next step, we removed all sentence pairs with scores below this threshold. Additionally, we also removed duplicate sentence pairs. The filtering step reduced the number of sentence pairs by about 15–20% (details in Table 3).

4.3 Manual corrections of parallel data

The translation quality of the parallel data obtained by the procedure described above is not guaranteed

[8] http://code.google.com/p/boilerpipe/

[9] http://bitextor.sourceforge.net/

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

languages (L1–L2)	dom	set	sentences	L1 tokens / vocabulary		L2 tokens / vocabulary	
English–French	*env*	dev	1,392	35,094	5,245	40,919	6,024
	env	test	2,000	49,778	6,252	58,166	7,252
	lab	dev	1,411	45,306	5,034	51,372	5,994
	lab	test	2,000	62,070	6,031	70,534	7,274
English–Greek	*env*	dev	1,000	26,507	5,790	23,980	3,980
	env	test	2,000	55,090	8,715	49,925	5,503
	lab	dev	506	14,169	3,509	13,201	2,453
	lab	test	2,000	58,429	7,466	54,372	4,559

Table 4: Development and test data set statistics.

in any sense. Tuning the procedure and focusing on high-quality translations is possible but leads to a trade-off between quality and quantity. For translation model training, high translation quality of the data is not as essential as for parameter tuning and testing. Bad phrase pairs can be removed from the translation tables based on their low translation probabilities. However, a development set containing sentence pairs which are not exact translations of each other might lead to sub-optimal values of model weights which would harm system performance. If such sentence pairs are used in the test set, the evaluation would clearly be very unreliable.

In order to create reliable development and test sets for each language pair and domain, we performed the following low-cost procedure. From the data obtained by the steps described in the previous section, we selected a random sample of 3,600 sentence pairs (2,700 for English–Greek in the Labour Legislation domain, for which no more data was available) and asked native speakers to check and correct them. The task consisted of:

1. checking that the sentence pairs belonged to the right domain,
2. checking that the sentences within a sentence pair were equivalent in terms of content,
3. checking translation quality and correcting (if needed) the sentence pairs.

The goal was to obtain at least 3,000 correct sentence pairs (2,000 test pairs and 1,000 development pairs) for each domain and language pair; thus the correctors did not have to correct every sentence pair. They were allowed to skip (remove) those sentence pairs which were misaligned. In addition, we asked them to remove those sentence pairs that were obviously from a very different domain (despite being correct translations). The number of corrected sentences obtained is shown in the last column of Table 3. As the final step, we took a random sample from the corrected sentence pairs and selected 2,000 pairs for the test set and left the remaining part for the development set.

During the correction phase, we made the following observations: 55% of sentence pairs were accurate translations, 35% of sentence pairs needed only minor corrections, 3–4% of sentence pairs would require major corrections (which was not necessary to do in most cases, as the accurate sentence pairs together with those requiring minor corrections were enough to reach our goal of at least 3,000 sentence pairs), 4–5% of sentence pairs were misaligned and would have had to be translated completely (which was not necessary in most cases), and 3–4% of sentence pairs were from a different domain. The correctors confirmed that the process was about 5–10 times faster than translating the sentences from scratch. Detailed statistics of the test and development sets obtained by the procedure described above are given in Table 4.

5 Experiments and results

The described approach was evaluated in eight different scenarios involving: two language pairs (English–Greek, English–French), both translation directions (to English and from English), and the two domains (Natural Environment, Labour Legislation), using the following automatic evaluation measures: WER, PER, and BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and METEOR (Banerjee and Lavie, 2005).

The baseline MT systems (denoted as *v0*) were evaluated using these test sets and results are shown in Table 5. The BLEU, METEOR, PER, and WER scores are percentages; WER and PER are error rates; OOV (out-of-vocabulary) is a ratio of unknown words, i.e. the occurrence of words which do not appear in the parallel training data and thus cannot be translated. The scores among different systems are not freely comparable but they give us some idea of how difficult translation is for particular languages or domains.

languages	dom	BLEU	NIST	METEOR	PER	WER	OOV
English → French	*env*	28.03	7.03	63.32	63.70	46.71	0.98
	lab	22.26	6.27	56.73	69.93	50.06	0.85
French → English	*env*	31.79	7.77	66.25	57.09	40.02	0.81
	lab	27.00	7.07	59.90	61.57	43.24	0.68
English → Greek	*env*	20.20	5.73	82.81	67.83	54.02	1.15
	lab	22.92	5.93	87.27	65.88	52.21	0.47
Greek → English	*env*	29.23	7.50	60.57	54.69	41.07	1.53
	lab	31.71	7.76	62.42	52.34	38.37	0.69

Table 5: Baseline MT system results (all scores except NIST are percentages).

The baseline MT system was trained solely on out-of-domain data (parallel, monolingual, and development data from Europarl). First, we exploited the in-domain development data and used it in the first modification of the baseline system (*v1*) instead of the out-of-domain (Europarl) data. In this case, the individual system models (translation tables, language model, etc.) remained the same, but their relative importance (optimal weights in the SMT log-linear framework) was different.

The in-domain monolingual data could be exploited in two ways: a) to join the general-domain data and the new in-domain data into one set, use it to train one language model and optimize its weight using MERT on the in-domain development data. b) to train a new separate language model from the new data and add it to the log-linear framework and let MERT optimize its weight together with other model weights. We tested both approaches. In system *v2* we followed the first option (retraining the language model on an enlarged data) and in system *v3* we followed the second option (training an additional language model and optimizing).

All evaluation results are presented in detail in Tables 6 to 9. Each table compares the performance of all systems (*v0*–*v3*) in one translation direction. The comparison between the scores of *v0* and *v1* tells us the importance of using in-domain data for parameter optimization. The improvement in terms of BLEU varies between 16% and 48% relative, which is quite substantial, especially given the fact that this modification requires several hundreds of sentence pairs only.

The comparison between *v1* and *v2* / *v3* shows the effect of using additional in-domain data for language modelling, which turned out not to be very substantial in most scenarios. With only one exception (see below), the BLEU scores improved by less than 1 point. This observation is not very surprising given the fact that the general-domain translation models were not enhanced in any way and thus the new in-domain language models had only limited room for improvement: the high OOV rates remained the same. After improving the translation models which (hopefully) will decrease the OOV rates, the language models might have a better chance of contributing to improved scores. The only exception was the translation from English to Greek for the Labour Legislation domain, for which the BLEU score increased massivelly from 28.79 to 33.43 points (Table 8). This is probably due to the richer morphology of Greek as the target language and the relatively low OOV rate on the Labour Legislation data; here, the performance improved even if the OOV rate did not change.

An analysis of the differences between the results of *v2* and *v3* could explain the difference between using in-domain monolingual data in one language model (together with general-domain data) vs. using two separate models (general--domain plus in-domain). However, due to the fact that the addition of in-domain monolingual data did not lead to any significant improvement in MT quality, the differences are not really measurable. It is likely that this situation will change after improving the translation models by adding in-domain parallel data.

6 Conclusion and future work

In this work we described our first steps towards domain adaptation of statistical machine translation based on data obtained by domain-focused web crawling. We evaluated four SMT systems in eight scenarios and tested the impact of two types of web-crawled language resources (in-domain parallel development data, in-domain monolingual training data) on the MT quality. In terms of automatic evaluation measures, the effect of using in-domain development data for parameter optimization in SMT is very substantial, in the range of 16–48% relative improvement. The impact of using in-domain monolingual data for language

sys	dom	BLEU / Δ%		NIST / Δ%		METEOR / Δ%		PER / Δ%		WER / Δ%	
v0	*env*	28.03	0.00	7.03	0.00	63.32	0.00	63.70	0.00	46.71	0.00
v1	*env*	35.81	27.76	8.10	15.22	68.44	8.09	53.78	-15.57	40.34	-13.64
v2	*env*	36.13	28.90	8.14	15.79	68.40	8.02	53.14	-16.58	40.07	-14.22
v3	*env*	36.32	29.58	8.19	16.50	68.50	8.18	52.82	-17.08	39.62	-15.18
v0	*lab*	22.26	0.00	6.27	0.00	56.73	0.00	69.93	0.00	50.06	0.00
v1	*lab*	30.84	38.54	7.42	18.34	62.94	10.95	57.99	-17.07	43.11	-13.88
v2	*lab*	30.18	35.58	7.31	16.59	62.86	10.81	59.05	-15.56	43.81	-12.49
v3	*lab*	30.12	35.31	7.28	16.11	62.88	10.84	59.48	-14.94	44.24	-11.63

Table 6: Evaluation results: English → French.

sys	dom	BLEU / Δ%		NIST / Δ%		METEOR / Δ%		PER / Δ%		WER / Δ%	
v0	*env*	31.79	0.00	7.77	0.00	66.25	0.00	57.09	0.00	40.02	0.00
v1	*env*	39.04	22.81	8.75	12.61	69.17	4.41	48.26	-15.47	34.56	-13.64
v2	*env*	39.27	23.53	8.77	12.87	69.26	4.54	48.16	-15.64	34.49	-13.82
v3	*env*	38.84	22.18	8.72	12.23	69.06	4.24	48.55	-14.96	34.71	-13.27
v0	*lab*	27.00	0.00	7.07	0.00	59.90	0.00	61.57	0.00	43.24	0.00
v1	*lab*	33.52	24.15	7.98	12.87	63.70	6.34	53.39	-13.29	38.42	-11.15
v2	*lab*	33.91	25.59	8.02	13.44	64.06	6.94	53.11	-13.74	38.22	-11.61
v3	*lab*	33.72	24.89	8.00	13.15	64.11	7.03	53.30	-13.43	38.31	-11.40

Table 7: Evaluation results: French → English.

modelling cannot be confirmed where a system has a high OOV rate, which can be minimized only by improving the coverage of the translation models. Our future work will focus on crawling more parallel data and enhancing the translation models. Results of tests of statistical significance will also be provided.

7 Acknowledgements

This work is supported by PANACEA, a 7th Framework Research Programme of the European Union, contract number 7FP-ITC-248064.

References

Ardö, Anders. 2005. Focused crawling in the ALVIS semantic search engine. In *Proceedings of the 2nd European Semantic Web Conference*, pages 19–20, Heraklion, Greece.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Banerjee, Pratyush, Jinhua Du, Baoli Li, Sudip Naskar, Andy Way, and Josef van Genabith. 2010. Combining multi-domain statistical machine translation models using automatic classifiers. In *AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, pages 141–150.

Baroni, M. and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th Language Resources and Evaluation Conference*, pages 1313–1316, Lisbon.

Doddington, George. 2002. 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 '02, pages 138–145, San Diego, California.

Eck, Matthias, Stephan Vogel, and Alex Waibel. 2004. Language model adaptation for statistical machine translation based on information retrieval. In *International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Esplà-Gomis, Miquel and Mikel L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathemathical Linguistics*, 93:77–86.

Finch, Andrew and Eiichiro Sumita. 2008. Dynamic model interpolation for statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 208–215, Columbus, Ohio, USA.

Hildebrand, Almut Silja, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, pages 133–142, Budapest, Hungary.

Hua, Wu, Wang Haifeng, and Liu Zhanyi. 2005. Alignment model adaptation for domain-specific word alignment. In *43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 467–474, Ann Arbor, Michigan, USA.

sys	dom	BLEU / Δ%		NIST / Δ%		METEOR / Δ%		PER / Δ%		WER / Δ%	
v0	*env*	20.20	0.00	5.73	0.00	82.81	0.00	67.83	0.00	54.02	0.00
v1	*env*	26.18	29.60	6.57	14.66	84.19	1.67	60.80	-10.36	49.10	-9.11
v2	*env*	26.50	31.19	6.63	15.71	84.35	1.86	60.65	-10.59	48.76	-9.74
v3	*env*	26.41	30.74	6.57	14.66	83.85	1.26	60.58	-10.69	48.99	-9.31
v0	*lab*	22.92	0.00	5.93	0.00	87.27	0.00	65.88	0.00	52.21	0.00
v1	*lab*	28.79	25.61	6.80	14.67	87.91	0.73	58.20	-11.66	46.43	-11.07
v2	*lab*	33.43	45.86	7.33	23.61	88.94	1.91	54.93	-16.62	43.77	-16.17
v3	*lab*	34.03	48.47	7.44	25.46	88.94	1.91	54.37	-17.47	43.25	-17.16

Table 8: Evaluation results: English → Greek.

sys	dom	BLEU / Δ%		NIST / Δ%		METEOR / Δ%		PER / Δ%		WER / Δ%	
v0	*env*	29.23	0.00	7.50	0.00	60.57	0.00	54.69	0.00	41.07	0.00
v1	*env*	34.16	16.87	8.01	6.80	64.98	7.28	51.15	-6.47	37.67	-8.28
v2	*env*	34.24	17.14	8.02	6.93	64.99	7.30	51.12	-6.53	37.65	-8.33
v3	*env*	34.15	16.83	8.01	6.80	64.75	6.90	51.09	-6.58	37.83	-7.89
v0	*lab*	31.71	0.00	7.76	0.00	62.42	0.00	52.34	0.00	38.37	0.00
v1	*lab*	37.55	18.42	8.28	6.70	67.36	7.91	49.02	-6.34	35.27	-8.08
v2	*lab*	38.00	19.84	8.36	7.73	67.73	8.51	48.45	-7.43	34.83	-9.23
v3	*lab*	37.70	18.89	8.32	7.22	67.40	7.98	48.76	-6.84	35.03	-8.70

Table 9: Evaluation results: Greek → English.

Koehn, Philipp and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Prague, Czech Republic.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Kohlschütter, Christian, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 441–450, New York.

Langlais, Philippe. 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology - Volume 14*, pages 1–7, Taipei, Taiwan.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31:477–504.

Nakov, Preslav. 2008. Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 147–150, Columbus, Ohio, USA.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *41st Annual Meeting on Association for Computational Linguistics*, ACL '03, pages 160–167, Sapporo, Japan.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, USA.

Penkale, Sergio, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, and Andy Way. 2010. MaTrEx: the DCU MT system for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 143–148, Uppsala, Sweden.

Stolcke, Andreas. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, pages 257–286, Denver, Colorado, USA.

Theobald, Martin, Jonathan Siddharth, and Andreas Paepcke. 2008. Spotsigs: robust and efficient near duplicate detection in large web collections. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 563–570, Singapore.

Wu, Hua and Haifeng Wang. 2004. Improving domain-specific word alignment with a general bilingual corpus. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, pages 262–271, Washington, DC.