Applications of Data Selection via Cross-Entropy Difference for Real-World Statistical Machine Translation

Amittai Axelrod, QingJun Li, William D. Lewis

Microsoft Research
Redmond, WA 98052, USA
amittai@alum.mit.edu
{v-qingjl,wilewis}@microsoft.com

Abstract

We broaden the application of data selection methods for domain adaptation to a larger number of languages, data, and decoders than shown in previous work, and explore comparable applications for both monolingual and bilingual cross-entropy difference methods. We compare domain adapted systems against very large general-purpose systems for the same languages, and do so without a bias to a particular direction. We present results against real-world general-purpose systems tuned on domain-specific data, which are substantially harder to beat than standard research baseline systems. We show better performance for nearly all domain adapted systems, despite the fact that the domain-adapted systems are trained on a fraction of the content of their general domain counterparts. The high performance of these methods suggest applicability to a wide variety of contexts, particularly in scenarios where only small supplies of unambiguously domain-specific data are available, yet it is believed that additional similar data is included in larger heterogenous-content general-domain corpora.

1. Introduction

The common wisdom in SMT is that "a lot of data is good" and "more data is better". This wisdom is backed up by evidence that scaling to ever larger data shows continued improvements in quality, even when one trains models over billions of n-grams [1]. Likewise, doubling or tripling the size of tuning data can show incremental improvements in quality as well [2]. Not all data is equal, however, and the kind of data one chooses depends crucially on the target domain. In a domain-specific setting, SMT benefits less from large amounts of general domain content; rather, it benefits from more content in the target domain, even if that content is appreciably smaller then the available pool of general content [3]. This fact has become more crucial as the community involved in the application of SMT has grown larger. The extended SMT community now includes an increasing number of multinational firms and public entities who wish to apply SMT to practical uses, such as automatically translating online knowledge bases, interacting with linguistically diverse customers over IM, translating large bodies of company-internal documentation for satellite offices, or even just broadening Web presence into new markets. For these new seats at the SMT table, data is still a gating factor for quality, but it is gated across another dimension: domain. For these SMT users, the rule really is not "more data is better", but rather its corollary, "more data *like my data* is better".

In this paper, we broaden the application of data selection methods for domain adaptation to a larger number of languages, data, and decoders than shown in previous work, and explore comparable applications for both monolingual [4] and bilingual [3] cross-entropy difference methods. The languages chosen for our study are typologically diverse, consisting of English, Spanish, Hebrew and Czech. A diverse sample of languages demonstrates that factors related to data sparsity, namely morphological complexity and structural divergence (*a la* [5]), are not significant factors in the successful application of the methods.

Further, we compare domain adapted systems against very large general purpose systems, whose data forms the supply of out-of-domain data we adapt from. Showing performance gains against such large systems ([3] constitutes prior work for Chinese-English) is a much harder baseline to beat than a simple out-of-the-box installation of a standard SMT toolkit. Our gains are made appreciably harder since we treat as one baseline a large general purpose system *tuned on target domain data*. For thoroughness, we also demonstrate resilience of the methodology to direction of translation, e.g., we not only apply the method to translating English $\rightarrow$ X but also to X $\rightarrow$ English, and to the decoder chosen, e.g., we use both phrase-based and tree-to-string decoders. In all cases, we demonstrate improvements in performance for domain-adapted systems over baselines that are trained on significantly larger supplies of data (10x more).

2. Task-Specific SMT

There has been much recent interest in methods for improving statistical machine translation systems targeted to a specific task or domain. The most common approach is that of

domain adaptation, whereby a system is trained on one kind of data, and then adjusted to apply to another. The adjustment can be as simple as retuning the model parameters on a task-specific dev set, such as [6]. Another common approach is to modifying the general-domain model using an in-domain model as a guide, or enhancing an in-domain model with portions of a general domain model, such as [7] among others.

We seek to accomplish the same goal as domain adaptation techniques, only by using the available data more effectively instead of modifying the model's contents. A *data selection* method is a procedure for ranking the elements of a pool of sentences using a relevance measure, and then keeping only the best-ranked ones. These data selection methods make binary decisions – keep or discard – but there are also soft-decision approaches, termed *instance weighting*.

Data selection methods have been used for some time in other NLP applications such as information retrieval (IR) (using tf-idf) and language modeling (using perplexity). One focus for those applications is mixture modeling, wherein data is selected to build sub-models, which are then weighted and combined into one larger model that is domain-specific [8]. These approaches were later combined by [9] and [10] to apply IR methods for build a translation mixture model using additional corpora. A different way of using all the available data yet highlighting its more relevant portions is to apply instance weighting. The main difference is that only one model is trained, rather than building multiple models and interpolating them against some held-out data. Experiments by [11] and [12] modified the n-gram counts from each sentence according to their relevance to the task at hand.

Moving away from mixture models, perplexity is commonly used as a selection criterion, such as by [13], to select additional training data for expanding a single in-domain language model. This method has the advantage of being extremely simple to apply: train a language model, score each additional sentence, and select the highest-ranked. This was applied to SMT by [14]. The main idea was repurposed by [4] to rank each additional sentence s by the *cross-entropy difference* between an in-domain language model and an LM trained on all of the additional data pool:

$$\underset{s \in POOL}{\text{argmin}} \quad H(s, LM_{IN}) - H(s, LM_{POOL})$$

The optimal selection threshold must be determined via grid search, but it is otherwise straightforward to apply. The cross-entropy difference criterion was first applied to the task of SMT by [3]. They also proposed a bilingual version of the criterion, consisting of the sum of the monolingual cross-entropy difference scores for two languages $L1$ and $L2$:

$$\begin{aligned}\underset{s \in POOL}{\text{argmin}} \quad & [H_{L1}(s, LM_{IN}) - H_{L1}(s, LM_{POOL})] \\ & +[H_{L2}(s, LM_{IN}) - H_{L2}(s, LM_{POOL})]\end{aligned}$$

Both the monolingual and bilingual versions have been used in recent SMT work, such as by [15] on Arabic-English and French-English, [16] for German-English and French-English systems, and in previous IWSLT evaluations for Chinese-English by [17] among others.

3. Effectiveness of Cross-Entropy Difference as a Data Selection Method

Our goal is to provide a more comprehensive survey of the impact of cross-entropy difference as a selection method for SMT. Cross-entropy difference has been shown to improve performance on domain-specific tasks, but to date the published work has focused on highly-constrained targets, such as IWSLT 2010 BTEC/DIALOG tasks and moderately-sized additional data (Europarl, UN corpora). The 2012 IWSLT TED talks are more realistic, as is the Gigaword corpus as a data pool. However, the TED talks exhibit great topical variety without a unifying domain. In this work we go further and provide experimental results on a broader, yet domain-specific, task and a much larger set of data to select from. As a result, we are in a position to evaluate the effectiveness of cross-entropy difference against a very large general-purpose statistical machine translation system, and examine the cases in which data selection may help. We also compare the relative effectiveness of the monolingual and bilingual versions of cross-entropy difference. We consequently built systems on three typologically diverse language pairs (Spanish/English, Czech/English, and Hebrew/English), in both translation directions. These corpora vary greatly in the amount of general bilingual training data available and the amount of bilingual in-domain data. Furthermore, we use two kinds of SMT systems to determine whether the system improvements depend on the flavor of SMT system used.

4. Experimental Setup

We used custom-built phrase-based and tree-to-string (T2S) systems for training the models for our engines. Our T2S decoder requires a source-side parser, and was used for all language pairs where the source had a parser: for all English → X pairs, as well as for Spanish → English. Lacking parsers for Czech and Hebrew, we used our custom built phrase-based decoder (functionally equivalent in many respects to the popular Moses phrase-based decoder [18]) to train the Czech → English and Hebrew → English systems.

For all English → X systems, we trained a 5-gram LM over all relevant monolingual data (the target side of the parallel corpus). Target side LMs for all X → English systems also used 5-gram LMs, trained over the target side of parallel data. For a subset of the systems in our study, we trained a second much larger 5-gram English language model over a much larger corpus of English language data (greater than 10 gigawords), including Web crawled content, licensed corpora (such as LDC's Gigaword), etc. We used Minimum Error Rate Training (MERT) [19] for tuning the lambda values for all systems, and results are reported in terms of BLEU score [20] on lowercased output with tokenized punctuation.

For the English → Spanish systems we trained a 5-gram LM, similar to that used for English, that is, one trained over Web crawled content, licensed corpora, and other sources. This LM was greater than 5 gigawords. For the equivalent English→Czech and English→ Hebrew systems, we built an additional 5-gram LM trained on the target side of the general purpose systems.

The bilingual general-purpose training data varied significantly between language pairs, reflecting the inconsistent availability of parallel resources for less common language pairs. As a result, we had 25 million sentences of parallel English-Spanish training data, 11 million sentences for Czech-English, and 3 million sentence pairs for Hebrew-English. In all cases these are significantly more data than has been made available for these language pairs in open MT evaluations, so this work addresses in part the question of how well the cross-entropy difference-based data selection methods scale.

Our target task is to translate travel-related information as might be written in guidebooks, online travel reviews, promotional materials, and the like. Note that this is significantly broader than much previous work in the travel domain, such as pre-2011 IWSLT tasks targeting conversational scenarios with a travel assistant. Our in-domain data for the Spanish-English language pair consisted of online travel review content, manually translated from English into Spanish (using Mechanical Turk), and a set of phrasebooks between English and Spanish. The total parallel in-domain content consisted of approximately 4 thousand sentences, which was strictly used for tuning and testing. For the monolingual selection methods, we used a corpus of online travel content in English, travel guidebooks, and travel-related phrases. This corpus consisted of approximately 600 thousand sentences.

For Czech-English and Hebrew-English we used translated travel guidebooks, consisting of 129k and 74k sentences (2.1m words and 1.2m words), respectively. The monolingual methods for these two language pairs, unlike Spanish-English, used the English side of the Czech-English and Hebrew-English guidebook (respectively). For these two language pairs we can therefore directly compare the monolingual and bilingual data selection methods. The held-out development and test sets for the Spanish-English systems consisted of crowdsourced human translations of data from a travel review website. For Czech-English and Hebrew-English, we used held-out portions of the same guidebooks used for the training data.

Because our baseline comparison is against a real-world SMT system, we used additional monolingual resources to train an output-side language model, and used it in lieu of an LM trained only on the output side of the parallel training corpus. We used the same LM for all X→English systems. The large monolingual LM ("All-mono" in the tables below) consistently yielded +0.75-3 BLEU over using only the output side of the bilingual training data. We are thus able to compare the performance of translation models trained on only a subset of the parallel data vs ones trained on all the data, without having to worry about the effect of the data selection process on LM coverage, as LM size and coverage has a substantial impact on SMT system performance.

In all cases, we built the following systems:

1. A baseline using all the available bilingual data to train the translation model, and all available monolingual data in the output language to train the language model. This system is tuned on a standard non-travel dev set (*e.g. WMT2010*), and represents a baseline of a very large scale SMT system with no adaptation.
2. Another baseline using all the available bilingual data to train the translation model, and all available monolingual data in the output language to train the language model. This baseline is tuned on the travel-specific devset for the language pair. Due to the size of the corpora involved, this may be considered a difficult baseline and is also the easiest way to build a domain-specific system using an existing general SMT system, since it does not require retraining.
3. An SMT system using only the top 10% of the bilingual training corpus to train the translation model, with the language model trained on the target side of this subset. The quantity of 10% was chosen empirically as generally representative of a well-performing adapted SMT system.
4. An SMT system using only the top 10% of the bilingual training corpus to train the translation model, but with a language model trained on all available monolingual data (like the baseline systems). This is more realistic than System #3 above, as it shows the effect of just reducing the size of the phrase table training corpus, but does not affect its ability to assemble fluent output.
5. A system with one translation model and one language model trained on the top 10%, as in System #3, but with the addition of a second language model using all the monolingual data.
6. A system with one translation model and one language model trained on the top 10%, as in System #3, but with the addition of a second translation model using all the bilingual data and a second language model using all the monolingual data. This is a general-purpose SMT system that has been augmented with a domain-specific phrase table and language model, and reflects what is achievable by considering all sources of training data for task-specific performance.

5. Results

5.1. Spanish↔English Language Pair

The English-Spanish language pair is the one with the most available general-coverage parallel data: 25 million

sentences. This is 20% larger than any previous cross-entropy difference experiment (*c.f.* 21m sentence pairs for English→French in [15]). This amount of data means the large-scale translation system is reasonably strong. For example, the baseline English→Spanish BLEU score on the WMT 2010 test set is 32.21, when tuned on the WMT 2010 dev set (see Table 1). However, this is also a language pair with an extremely limited amount of parallel travel-specific data: practically none, as there is not enough to train even a language model on. In this situation, we assembled all available monolingual English travel data (consisting of the English half of bilingual travel data for other language pairs) and used it exclusively to select relevant training data from the large Spanish-English corpus.

The English↔Spanish systems were tuned on 2,930 travel review sentences, and tested on 776 sentences from the same source. We used an additional 992 travel-related sentences translated from online hotel reviews as a second test set. Of interest also is the degradation in performance of a travel-tuned system on non-travel data, so we evaluated all the systems on the WMT2010 test set. Results for English→Spanish are in Table 1, and for Spanish→English are in Table 2.

Table 1 shows that by augmenting the baseline system with the translation model and language model trained on the top 10% of the training data, it is possible to gain an extra +0.3 BLEU points on the travel task, an extra +0.6 BLEU on the hotel reviews, while only losing -0.2 on the WMT task compared to just retuning the baseline system on the travel devset. Depending on the application, this may be a worthwhile tradeoff. However – and as expected – overall performance on the general WMT2010 task decreases by over a BLEU point when tuning on the travel domain. This must be taken into consideration when deciding how to use existing SMT systems for additional tasks.

The results in Table 2 are similar in story; the main difference is that the impact of corpus size for language model training is more apparent because the output language is English. Using all monolingual data instead of just the bilingual corpus to train the LM adds at least 3 BLEU points to the score of all the systems that use it; this is why we use the large LM for all but one of our experimental SMT systems.

5.2. Czech↔English Language Pair

For the Czech↔English translation pair we have less than half as much parallel general-domain text (11m sentences) than the Spanish↔English pair, however, there is substantially more bilingual in-domain text. We are therefore able to compare the effectiveness of the monolingual vs bilingual selection methods for both translation directions. For the monolingual methods we build an LM on the English half of the travel data, and for the bilingual selection method we build language models on each side and apply them as per the equation in Section 2. The un-adapted baseline system is tuned on WMT dev2010, which is 4,807 sentences in size. The travel-adapted systems were tuned on 1,984 sentences of guidebook data, and the held-out test set consists of 4,844 sentences from the same guidebook. These datasets are large enough to provide stable and representative results.

We first examine results for the English → Czech direction, tabulated in Table 3. Tuning the baseline system on travel-specific data improved performance by +0.4 on the guidebook test set, but caused a loss of -0.5 on the WMT test set. When comparing against the domain-tuned baseline, we see that the models built on data selected via the monolingual cross-entropy method always decrease performance, if only slightly. The systems trained on data selected via the bilingual criterion do slightly better, but could be described as being at best equal to the baseline on the guidebook data, but are even worse on the WMT test set. We therefore have a case where cross-entropy difference as a data selection method does not outperform simply retuning an existing system on a dev set pertaining to the new target task.

Table 4 contains results from experiments in the other direction, from Czech → English. As before, the retuned baseline system gains +1.5 on the guidebook data, but loses -2 on the WMT. The data selection results, however, differ markedly from the other translation direction, even though the selection criteria are exactly the same. Using the monolingually-selected systems we can see that using the LM trained on the selected data is slightly harmful, but that the large language model is surprisingly powerful, making a +4 BLEU impact. The selected translation mode is good for a +2 BLEU improvement on its own, and using all the models together yields a +2.8 improvement over the retuned baseline on the guidebook data, at a cost of -1.4 to the WMT test set performance. The bilingually selected methods are consistently better, but only marginally so (+0.1 BLEU).

Thus data selection methods provide substantial improvements when translating Czech → English, and none from English → Czech. Two differences between the systems are that the former is a phrasal MT system, and the latter is a treelet translation system. Furthermore, the output language model is significantly better when translating into English than into Czech, simply due to the differing amounts of LM training data.

5.3. Hebrew↔English Language Pair

Our Hebrew↔English translation pair has the least amount of parallel training data of the ones we tested, but still has 3 million sentences, making it larger than the Europarl corpus which is a standard for European languages. The baseline large-scale system was tuned on 2,000 sentences extracted

Table 1: *English to Spanish*

Model	Phrase Table 1	TM 2	LM 1	LM 2	Travel Reviews	Hotel Reviews	WMT 2010
Baseline	All	–	All-mono	–	33.27	28.19	31.00
Baseline (WMT2010)	All	–	All-mono	–	32.28	**29.09**	**32.21**
Top 10% TM, All-mono LM	Top 10%	–	All-mono	–	32.78	28.09	28.07
Top 10% only	Top 10%	–	Top 10%	–	32.61	27.25	25.60
+All-mono LM	Top 10%	–	All-mono	Top 10%	33.12	28.18	28.19
+ All TM	Top 10%	All	All-mono	Top 10%	**33.55**	28.80	30.81

Table 2: *Spanish to English*

Model	Phrase Table 1	TM 2	LM 1	LM 2	Travel Reviews	Hotel Reviews	WMT 2010
Baseline	All	–	All-mono	–	39.43	32.79	31.38
Baseline (WMT2010)	All	–	All-mono	–	38.71	32.03	**32.11**
Top 10% only	Top 10%	–	Top 10%	–	37.18	30.04	26.48
+All-mono LM	Top 10%	–	All-mono	Top 10%	39.49	32.38	29.57
+All TM	Top 10%	All	All-mono	Top 10%	**40.00**	**33.28**	31.05

Table 3: *English to Czech*

Model	Phrase Table 1	TM 2	LM 1	LM 2	Guidebook	WMT 2010
Baseline	All	–	All-mono	–	27.73	15.03
Baseline WMT2010	All	–	All-mono	–	*27.33*	**15.59**
Monolingual Top 10% only	Top 10%	–	Top 10%	–	24.80	12.63
Monolingual Top 10% TM, All-mono LM	Top 10%	–	All-mono	–	27.84	13.95
+ Top 10%LM	Top 10%	–	All-mono	Top 10%	27.69	13.59
+ All TM	Top 10%	All	All-mono	Top 10%	27.43	14.25
Bilingual Top 10% only	Top 10%	–	Top 10%	–	24.92	12.52
Bilingual Top 10% TM only, All-mono LM	Top 10%	–	All-mono	–	27.68	13.67
+ Top 10% LM	Top 10%	–	All-mono	Top 10%	27.77	13.48
+ All TM	Top 10%	All	All-mono	Top 10%	**27.80**	*14.88*

Table 4: *Czech to English*

Model	Phrase Table 1	TM 2	LM 1	LM 2	Guidebook	WMT 2010
Baseline	All	–	All-mono	–	34.06	21.83
Baseline (WMT2010)	All	–	All-mono	–	32.52	**23.88**
Monolingual Top 10% only	Top 10%	–	Top 10%	–	30.48	15.86
Monolingual Top 10% TM, All-mono LM	Top 10%	–	All-mono	–	34.64	19.46
+ Top 10% LM	Top 10%	–	All-mono	Top 10%	34.32	19.36
+ All TM	Top 10%	All	All-mono	Top 10%	35.36	22.40
Bilingual Top 10% only	Top 10%	–	Top 10%	–	30.64	15.90
Bilingual Top 10% TM, All-mono LM	Top 10%	–	All-mono	–	34.66	19.51
+ Top 10% LM	Top 10%	–	All-mono	Top 10%	34.55	19.38
+ All TM	Top 10%	All	All-mono	Top 10%	**35.48**	22.15

from the results of web queries. The travel domain data, like for Czech↔English, consists of travel guidebooks. We held out 1,979 sentences as a development set, plus an additional 4,764 sentences as a stable test set. We also report results on the WMT 2009 test set, so as to provide a comparison with other published work in SMT.

The results for translating from English→Hebrew are shown in Table 5. Retuning the baseline general-domain system on the travel dev set increases the BLEU score on the guidebook test set by +0.4, at a cost of -0.3 on the WMT 2009 set. There is not much difference in the results from selecting the best 10% of the general training corpus with the monolingual vs bilingual cross-entropy difference. In both cases, adding an LM trained on the selected data does no better than just using the largest LM possible. However, just using the most relevant data for a translation model provides a slight improvement (+0.3), and augmenting the baseline system with models trained on just the best selected data provide a total improvement of +1 BLEU on the guidebook test set. The only difference between the monolingual and bilingual versions of the selection criterion is that the best monolingually-selected system loses only -0.1 BLEU on the unrelated WMT 2009 test set, compared to -0.7 with the bilingually-selected equivalent.

Results for data selection for Hebrew→English systems can be found in Table 5. Retuning the existing large-scale baseline system provides a +0.4 increase on the guidebook test set, and a +0.1 improvement on the WMT set. The latter is slightly unexpected. However, using cross-entropy difference to augment the SMT system provides a total improvement of almost +1 BLEU.

In general, the systems selected by monolingual cross-entropy difference do the same as their counterparts picked using bilingual cross-entropy difference, if not marginally better. Unlike in the previous translation direction, replacing the general-domain phrase table with one built on the most-relevant 10% of the training data generally made things slightly worse. Only augmenting the general system with the models trained on the selected subsets improved performance over the retuned baseline. As before, the gain of +0.7 BLEU on the guidebook test set was offset by a loss of -0.2 to -0.5 on the WMT 2009 test set.

6. Analysis

Generally, the difference between monolingual-on-English side and bilingual cross-entropy difference was minor. This is in contrast to prior work on Chinese→English, which suggested that the bilingual method was notably better [3]. One key difference between that work and this one is that they tested monolingual methods on the input side, namely Chinese. In this work the monolingual method was always computed using the English language, regardless of whether it was input or output. It may simply be that the monolingual cross-entropy difference score is sufficient, if the language used for the selection criterion is capable of being well-represented by an n-gram model by virtue of having simpler morphology or lesser long-range dependencies than the other member of the language pair. When it is unclear which of the two languages is better suited, then the bilingual cross-entropy method is a safe choice, as it provides generally the same effectiveness and does not seem to do any harm. That said, the experiments on Spanish↔English confirm prior work that bilingual in-domain data is not strictly necessary to adapt an SMT system to a target task.

Only one translation direction English↔Czech showed no need for data selection. In that particular case, the same improvement could be obtained by simply retuning the existing general-purpose system. However, Czech is the most morphologically complex of the languages used in this work and one could argue that it therefore suffers more from n-gram sparsity than other languages when trying to build a translation or language model on a corpus of a specific size. That the average English↔Czech system score was 7 BLEU points lower than the reverse translation direction points to the difficulty of translating into Czech. Perhaps the optimal number of sentences to select is substantially larger than for other language pairs, and so that 10% of the data could produce a system equally good as a system on the full data simply means if 20 or 30% of the data were selected then one might see a significant improvement beyond that baseline.

The overall scores for translating Hebrew↔English were the lowest, presumably due to morphological complexity coupled with the least amount of training data. Nonetheless, the gains from domain adaptation via data selection were still large in both directions. The systems trained on data selected with bilingual cross-entropy difference performed similarly on the guidebook test set as the ones trained on monolingually-selected data. However, the bilingually-selected systems performed slightly worse on the WMT 2009 test set, raising the same question as English↔Czech: how much of a morphologically rich language can be usefully captured by an n-gram language model trained on a small in-domain corpus?

Interestingly, translating into English was always improved using data selection methods. This is somewhat counterintuitive, as the larger output-side language model might be assumed to mask changes to the other components of the SMT system, much as a larger language model is assumed to always improve translation output. Furthermore, reducing the size of the language model always hurt significantly, and the best systems always included the largest LM. This may indicate that it is less important to adapt the language model than it is to provide more domain-accurate phrase tables.

In most cases, the performance improvement on the travel task of a task-specific SMT system was greater than the performance loss on the regular test set (*e.g.* WMT test

Table 5: *English to Hebrew*

Model	Phrase Table 1	TM 2	LM 1	LM 2	Guidebook	WMT 2009
Baseline	All	–	All-mono	–	12.45	14.53
Baseline ReqLog	All	–	All-mono	–	12.04	**14.88**
Monolingual Top 10%	Top 10%	–	Top 10%	–	10.37	10.17
Monolingual Top 10% TM only	Top 10%	–	All-mono	–	12.79	11.75
+All-mono LM	Top 10%	–	All-mono	Top 10%	12.77	11.57
+ All TM	Top 10%	All	All-mono	Top 10%	**13.46**	14.43
Bilingual Top 10%	Top 10%	–	Top 10%	–	10.33	10.01
Bilingual Top 10% TM only	Top 10%	–	All-mono	–	12.88	11.55
+All-mono LM	Top 10%	–	All-mono	Top 10%	12.80	11.66
+ All TM	Top 10%	All	All-mono	Top 10%	**13.49**	13.84

Table 6: *Hebrew to English*

Model	Phrase Table 1	TM 2	LM 1	LM 2	Guidebook	WMT 2009
Baseline	All	–	All-mono	–	18.58	**25.18**
Baseline ReqLog	All	–	All-mono	–	18.18	25.03
Monolingual Top 10%	Top 10%	–	Top 10%	–	16.47	16.08
Monolingual Top 10% TM only	Top 10%	–	All-mono	–	18.13	19.36
+All-mono LM	Top 10%	–	All-mono	Top 10%	18.17	19.54
+ All TM	Top 10%	All	All-mono	Top 10%	**19.12**	24.92
Bilingual Top 10%	Top 10%	–	Top 10%	–	16.46	16.15
Bilingual Top 10% TM only	Top 10%	–	All-mono	–	18.09	19.16
+All-mono LM	Top 10%	–	All-mono	Top 10%	18.20	18.85
+ All TM	Top 10%	All	All-mono	Top 10%	**19.05**	24.77

2010). This implies that the trade-offs between performance on two distinct targets are not unbounded: one rarely loses more than one gets. Thus one may make an informed decision as to whether domain adaptation is worth while by comparing against acceptable drops in performance on other tasks of interest.

Finally, despite half of the translation systems being built using phrase-based SMT and the other half with syntactic/treelet systems, this does not seem to have an obvious impact on the appropriateness of data selection methods for improving in-domain performance.

7. Conclusions

We have presented a broader survey of tailoring a general translation system to a target task by selecting a subset of the training data using cross-entropy difference. We performed experiments in both translation directions for three language pairs. These language pairs exhibit varying levels of morphological complexity, amounts of parallel general-purpose data, and amounts of parallel in-domain data. We systematically compared methods of using the selected training data against real-world baselines consisting of very large general-purpose SMT systems using all available additional monolingual resources for language models, and show gains over these baselines of +0.3/1.3 BLEU for Spanish↔English, +0.5/3.0 for Czech↔English, and +0.7/1.4 for Hebrew↔English. These results confirm all prior work showing that only a fraction of general-purpose data is needed for a task-specific SMT system of at least equivalent performance on the domain of interest. We have also shown how domain adaptation adversely affects performance on non-domain-specific tasks, but the results also indicate that the loss in performance on a general task is often less than the improvement on the domain of interest, both quantifying and arguably justifying the tradeoff.

8. Acknowledgements

We gratefully acknowledge the assistance of Marco Chierotti in acquiring the crowdsourced translations of the travel domain data.

9. References

[1] Brants, T., Popat, A., Xu, P., Och, F., Dean, J. “Large Language Models in Machine Translation.” EMNLP (Empirical Methods in Natural Language Processing). 2007.

[2] Koehn, P,. and Haddow, B. “Towards Effective Use of Training Data in Statistical Machine Translation.” WMT (Workshop on Statistical Machine Translation). 2012.

[3] Axelrod, A., He, X., and Gao, J. “Domain Adaptation via Pseudo In-Domain Data Selection”. EMNLP (Empirical Methods in Natural Language Processing). 2011.

[4] Moore, R. C. and Lewis, W. “Intelligent Selection of Language Model Training Data”. ACL (Association for Computational Linguistics). 2010.

[5] Dorr, B. “Machine Translation Divergences: A Formal Description and Proposed Solution.” ACL (Association for Computational Linguistics). 1994.

[6] Li, M., Zhao, Y., Zhang, D., and Zhou, M. “Adaptive Development Data Selection for Log-linear Model in Statistical Machine Translation”. COLING (International Conference on Computational Linguistics). 2010.

[7] Bisazza, A., Ruiz, N., Federico, M. “Fill-Up versus Interpolation Methods for Phrase-Based SMT Adaptation”. WMT (Workshop on Statistical Machine Translation). 2011.

[8] Iyer, R., Ostendorf, M., and Gish, H. “Using Out-of-Domain Data to Improve In-Domain Language Models”. IEEE Signal Processing Letters. 4(8):221-223. 1997.

[9] Lu, Y., Huang, J., and Liu, Q. “Improving Statistical Machine Translation Performance by Training Data Selection and Optimization.” EMNLP (Empirical Methods in Natural Language Processing). 2007.

[10] Foster, G., and Kuhn, R. “Mixture-Model Adaptation for SMT”. WMT (Workshop on Statistical Machine Translation). 2007.

[11] Matsoukas, S., Rosti, A.-V., and Zhang, B. “Discriminative Corpus Weight Estimation for Machine Translation.” EMNLP (Empirical Methods in Natural Language Processing). 2009.

[12] Foster, G., Goutte, C., and Kuhn, R. “Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation.” EMNLP (Empirical Methods in Natural Language Processing). 2010.

[13] Gao, J., Goodman, J., Li, M., and Lee, K.-F. “Toward a Unified Approach to Statistical Language Modeling for Chinese”. ACM Transactions On Asian Language Information Processing. 1(1):333. 2002.

[14] Yasuda, K., Zhang, R., Yamamoto, H., and Sumita, E. “Method of Selecting Training Data to Build a Compact and Efficient Translation Model.” IJCNLP (International Joint Conference on Natural Language Processing). 2008.

[15] Mansour, S., Wuebker, J., and Ney, H. “Combining Translation and Language Model Scoring for Domain-Specific Data Filtering.” IWSLT (International Workshop on Spoken Language Translation). 2011.

[16] Banerjee, P., Kumar, S., Roturier, J., Way, A., van Genabith, J. “Translation Quality-Based Supplementary Data Selection by Incremental Update of Translation Models”. COLING (International Conference on Computational Linguistics). 2012.

[17] He, X., Axelrod, A., Deng, L., Acero, A., Hwang, M.-Y., Nguyen, A., Wang, A., and Huang, X. “The MSR System for IWSLT 2011 Evaluation”. IWSLT (International Workshop on Spoken Language Translation). 2011.

[18] Koehn, P., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Moran, C., Dyer, C., Constantin, A., and Herbst, E. “Moses: Open Source Toolkit for Statistical Machine Translation”. ACL (Association for Computational Linguistics) Interactive Poster and Demonstration Sessions. 2007.

[19] Och, F. “Minimum Error Rate Training in Statistical Machine Translation.” ACL (Association for Computational Linguistics). 2003.

[20] Papineni, K., Roukos, S., Ward, T., and Zhu, W. “BLEU: a Method for Automatic Evaluation of Machine Translation”. ACL (Association for Computational Linguistics). 2002.

[21] Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. “Does More Data Always Yield Better Translations?” EACL (European Association for Computational Linguistics). 2012.

[22] Federico, M. “Language Model Adaptation through Topic Decomposition and MDA Estimation.” ICASSP (International Conference on Acoustics, Speech, and Signal Processing). 2002.

[23] Quirk, C., and Menezes, A. “Dependency Treelet Translation: The Convergence of Statistical and Example-Based Machine Translation?” Machine Translation. 20:43-65. 2006.

[24] Quirk, C. and Moore, R. “Faster Beam-Search Decoding for Phrasal Statistical Machine Translation”. Machine Translation Summit XI. 2007.

[25] He, X., and Deng, L. “Robust Speech Translation by Domain Adaptation.” Interspeech. 2011.