Language and Translation Model Adaptation using Comparable Corpora

Matthew Snover and **Bonnie Dorr**
Laboratory for Computational Linguistics
and Information Processing
Institute for Advanced Computer Studies
Department of Computer Science
University of Maryland
College Park, MD 20742
{snover, bonnie}@umiacs.umd.edu

Richard Schwartz
BBN Technologies
10 Moulton Street
Cambridge, MA 02138, USA
schwartz@bbn.com

Abstract

Traditionally, statistical machine translation systems have relied on parallel bi-lingual data to train a translation model. While bi-lingual parallel data are expensive to generate, monolingual data are relatively common. Yet monolingual data have been under-utilized, having been used primarily for training a language model in the target language. This paper describes a novel method for utilizing monolingual target data to improve the performance of a statistical machine translation system on news stories. The method exploits the existence of comparable text—multiple texts in the target language that discuss the same or similar stories as found in the source language document. For every source document that is to be translated, a large monolingual data set in the target language is searched for documents that might be comparable to the source documents. These documents are then used to adapt the MT system to increase the probability of generating texts that resemble the comparable document. Experimental results obtained by adapting both the language and translation models show substantial gains over the baseline system.

1 Introduction

While the amount of parallel data available to train a statistical machine translation system is sharply limited, vast amounts of monolingual data are generally available, especially when translating to languages such as English. Yet monolingual data are generally only used to train the language model of the translation system. Previous work (Fung and Yee, 1998; Rapp, 1999) has sought to learn new translations for words by looking at comparable, but not parallel, corpora in multiple languages and analyzing the co-occurrence of words, resulting in the generation of new word-to-word translations.

More recently, Resnik and Smith (2003) and Munteanu and Marcu (2005) have exploited monolingual data in both the source and target languages to find document or sentence pairs that appear to be parallel. This newly discovered bilingual data can then be used as additional training data for the translation system. Such methods generally have a very low yield leaving vast amounts of data that is only used for language modeling.

These methods rely upon comparable corpora, that is, multiple corpora that are of the same general genre. In addition to this, documents can be comparable—two documents that are both on the same event or topic. Comparable documents occur because of the repetition of information across languages, and in the case of news data, on the fact that stories reported in one language are often reported in another language. In cases where no direct translation can be found for a source document, it is often possible to find documents in the target language that are on the same story, or even on a related story, either in subject matter or historically. Such documents can be classified as comparable to the original source document. Phrases within this comparable document are likely to be translations of phrases in the source document, even if the documents themselves are not parallel.

Figure 1 shows an excerpt of the reference translation of an Arabic document, and figure 2 shows a

Cameras are flashing and reporters are following up, for Hollywood star Angelina Jolie is finally talking to the public after a one-month stay in India, but not as a movie star. The Hollywood actress, goodwill ambassador of the United Nations high commissioner for refugees, met with the Indian minister of state for external affairs, Anand Sharma, here today, Sunday, to discuss issues of refugees and children. **...** Jolie, accompanied by her five-year-old son, Maddox, visited the refugee camps that are run by the Khalsa Diwan Society for social services and the high commissioner for refugees Saturday afternoon after she arrived in Delhi. Jolie has been in India since October 5th shooting the movie "A Mighty Heart," which is based on the life of Wall Street Journal correspondent Daniel Pearl, who was kidnapped and killed in Pakistan. Jolie plays the role of Pearl's wife, Mariane.

Figure 1: Excerpt of Example Reference Translation of an Arabic Source Document

comparable passage.[1] In this case, the two new stories are not translations of each other and were not reported at the same time—the comparable passage being an older news story—but both discuss actress Angelina Jolie's visit to India. Many phrases and words are shared between the two, including: the name of the movie, the name and relationship of the actress' character, the name and age of her son and many others. Such a pairing is extremely comparable, although even less related document pairs could easily be considered comparable.

We seek to take advantage of these comparable documents to inform the translation of the source document. This can be done by augmenting the major components of the statistical translation system: the Language Model and the Translation Model. This work is in the same tradition as Kim and Khudanpur (2003), Zhao et al. (2004), and Kim (2005). Kim (2005) used large amounts of comparable data to adapt language models on a document-by-document basis, while Zhao et al. (2004) used comparable data to perform sentence level adaptation of the language model. These adapted language models were shown to improve performance

[1]This is an actual source document from the tuning set used in our experiments, and the first of a number of similar passages found by the comparable text selection system described in section 2.

Actress Angelina Jolie hopped onto a crowded Mumbai commuter train Monday to film a scene for a movie about slain journalist Daniel Pearl, who lived and worked in India's financial and entertainment capital. Hollywood actor Dan Futterman portrays Pearl and Jolie plays his wife Mariane in the "A Mighty Heart" co-produced by Plan B, a production company founded by Brad Pitt and his ex-wife, actress Jennifer Aniston. Jolie and Pitt, accompanied by their three children – Maddox, 5, 18-month-old Zahara and 5-month-old Shiloh Nouvel – arrived in Mumbai on Saturday from the western Indian city Pune where they were shooting the movie for nearly a month. **...**

Figure 2: Excerpt of Example Comparable Document

for both automatic speech recognition as well as machine translation.

In addition to language model adaptation we also modify the translation model, adding additional translation rules that enable the translation of new words and phrases in both the source and target languages, as well as increasing the probability of existing translation rules. Translation adaptation using the translation system's own output, known as Self-Training (Ueffing, 2006) has previously shown gains by augmenting the translation model with additional translation rules. In that approach however, the translation model was augmented using parallel data, rather than comparable data, by interpolating a translation model trained using the system output with the original translation model.

Translation model adaptation using comparable out-of-domain parallel data, rather than monolingual data was shown by Hildebrand et al. (2005) to yield significant gains over a baseline system. The translation model was adapted by selecting comparable sentences from parallel corpora for each of the sentences to be translated. In addition to selecting out-of-domain data to adapt the translation model, comparable data selection techniques have been used to select and weight portions of the existing training data for the translation model to improve translation performance (Lu et al., 2007).

The research presented in this paper utilizes a different approach to translation model adaptation using comparable monolingual text rather than parallel text, exploiting data that would otherwise be unused

for estimating the translation model. In addition, this data also informs the translation system by interpolating the original language model with a new language model trained from the same comparable documents.

We discuss the selection of comparable text for model adaptation in section 2. In sections 3.1 and 3.2, we describe the model adaptation for the language model and translation model, respectively. Experimental results describing the application of model adaptation to a hierarchical Arabic-to-English MT system are presented in section 4. Finally we draw conclusions in sections 5.

2 Comparable Text Selection

Comparable text is selected for every source document from a large monolingual corpus in the target language. In practice, one could search the World Wide Web for documents that are comparable to a set of source documents, but this approach presents problems for ensuring the quality of the retrieved documents. The experiments in this paper use comparable text selected from a collection of English news texts. Because these texts are all fluent English, and of comparable genre to the test set, they are also used for training the standard language model training.

The problem of selecting comparable text has been widely studied in the information retrieval community and cross-lingual information retrieval (CLIR) (Oard and Dorr, 1998; Levow et al., 2005) has been largely successful at the task of selecting comparable or relevant documents in one language given a query in another language. We use CLIR to select a ranked list of documents in our target language, English in the experiments described in this paper, for each source document, designated as the query in the CLIR framework, that we wish to translate.

The CLIR problem can be framed probabilistically as: Given a query Q, find a document D that maximizes the equation $\Pr(D \text{ is rel}|Q)$. This equation can be expanded using Bayes' Law as shown in equation 1. The prior probability of a document being relevant can be viewed as uniform, and thus in this work, we assume $\Pr(D \text{ is rel})$ is a constant.[2] The $\Pr(Q)$ is constant across all documents. Therefore finding a document to maximize $\Pr(D \text{ is rel}|Q)$ is equivalent to finding a document that maximizes $\Pr(Q|D \text{ is rel})$.

$$\Pr(D \text{ is rel}|Q) = \frac{\Pr(D \text{ is rel}) \Pr(Q|D \text{ is rel})}{\Pr(Q)} \quad (1)$$

A method of calculating the probability of a query given a document was proposed by (Xu et al., 2001)[3] and is shown in Equation 2. In this formulation, each foreign word, f, in the query is generated from the foreign vocabulary with probability α and from the English document with probability $1 - \alpha$, where α is a constant.[4] The probability of f being generated by the general foreign vocabulary, F, is $\Pr(f|F) = \mathit{freq}(f, F)/|F|$, the frequency of the word f in the vocabulary divided by the size of the vocabulary. The probability of the word being generated by the English document is the sum of the probabilities of it being generated by each English word, e, in the document which is the frequency of the English word in the document, ($\Pr(e|D) = \mathit{freq}(e, D)/|D|$) multiplied by the probability of the translation of the English word to the foreign word, $\Pr(f|e)$.

$$\Pr(Q|D) = \prod_{f \in Q} (\alpha \Pr(f|F) + (1 - \alpha) \sum_{e} \Pr(e|D) \Pr(f|e)) \quad (2)$$

This formulation favors longer English documents over shorter English documents. In addition, many documents cover multiple stories and topics. For the purposes of adaptation, shorter, fully comparable documents are preferred to longer, only partially comparable documents. We modify the CLIR system by taking the 1000 highest ranked target language documents found by the CLIR system for each source document, and dividing them into overlapping passages of approximately 300 words.[5] Sen-

[2] In fact, it can be beneficial to use features of the document to estimate $\Pr(D \text{ is rel})$ (Miller and Schwartz, 1998) but we have not explored that here.

[3] Xu et al. (2001) formulated this for the selection of foreign documents given an English query. We reverse this to select English documents given a foreign query.

[4] As in Xu et al. (2001), a value of 0.3 was used for α.

[5] The length of 300 was chosen as this was approximately the same length as the source documents.

tence boundaries are preserved when creating passages, insuring that the text is fluent within each passage. These passages are then scored again by the CLIR system, resulting in a list of passages of about 300 words each for each source document. Finally, we select the top N passages to be used for adaptation.

The N passages selected by this method are not guaranteed to be comparable and are often largely unrelated to the story or topic in the source document. We shall refer to the set of passages selected by the CLIR system as the *bias text* to differentiate it from comparable text, as the adaptation methods will use this text to *bias* the MT system so that its output will be more similar to the bias text.

While we have not conducted experiments using other CLIR systems, the adaptation methods presented in this paper could be applied without modification using another CLIR system, as the adaptation method treats the CLIR system as a black box. With the exception of running a second pass of CLIR, we use the algorithm of Xu et al. (2001) without any significant modification, including the use of a stop word list for both the English and foreign texts. The parameters for $\Pr(f|F)$ and $\Pr(f|e)$ were estimated using the same parallel data that our translation system was trained on.

The bias text selected for a source document is used to adapt the language model (described in section 3.1) and the translation model (described in section 3.2) when translating that source document.

3 Model Adaptation

We use the same bias text to adapt both the language model and the translation model. For language model adaptation, we increase the probability of the word sequences in the bias text, and for translation model adaptation we use additional phrasal translation rules. The adaptations can be done independently and while they can augment each other when used together, this is not required. It is not necessary to use the same number of passages for both forms of adaptation, although doing so makes it more likely both that the English side of the new translation rule will be assigned a high probability by the adapted language model, and that the translation model produces the English text to which the language model has been adapted. Bias text that is used by one adaptation but not the other will receive no special treatment by the other model. This could result in new translation rules that produce text to which the language assigns low probability, or it could result in the language model being able to assign a high probability to a good English translation that cannot be produced by the translation model due to a lack of necessary translation rules.

While both adaptation methods are integrated into a hierarchical translation model (Chiang, 2005), they are largely implementation independent. Language model adaptation could be integrated into any statistical machine translation that uses a language model over words, while translation model adaptation could be added to any statistical machine translation that can utilize phrasal translation rules.

3.1 Language Model Adaptation

For every source document, we estimate a new language model, the *bias language model*, from the corresponding bias text. Since this bias text is short, the corresponding bias language model is small and specific, giving high probabilities to those phrases that occur in the bias text. The bias language model is interpolated with the *generic language model* that would otherwise be used for translation if no LM adaptation was used. The new bias language model is of the same order as the generic language model, so that if a trigram language model is used for the MT decoding, then the biased language model will also be a trigram language model. The bias language model is created using the same settings as the generic language model. In our particular implementation however, the generic language model uses Kneser-Ney smoothing, while the biased language model uses Witten-Bell smoothing due to implementation limitations. In principle the biased language model can be smoothed in the same manner as the generic language model.

We interpolate the bias language model and the generic language model as shown in equation 3, where $\Pr_g$ and $\Pr_b$ are the probabilities from the generic language model and the bias language model, respectively. A constant interpolation weight, λ is used to weight the two probabilities for all documents. While a value for λ could be chosen that minimizes perplexity on a tuning set, in a

similar fashion to Kim (2005), it is unclear that such a weight would be ideal when the interpolated language model is used as part of a statistical translation system. In practice we have observed that weights other than one that minimizes perplexity, typically a lower weight, can yield better translation results on the tuning set.

$$\Pr(e) = (1 - \lambda) \Pr_{g}(e) + \lambda \Pr_{b}(e) \quad (3)$$

The resulting interpolated language model is then used in place of the generic language model in the translation process, increasing the probability that the translation output will resemble the bias text. It is important to note that, unlike the translation model adaptation described in section 3.2, no new information is added to the system with language model adaptation. Because the bias text is extracted from the same monolingual corpus that the generic language model was estimated from, all of the word sequences used for training the bias language model were also used for training the generic language model. Language model adaptation only increases the weight of the portion of the language model data that was selected as comparable.

3.2 Translation Model Adaptation

It is frequently the case in machine translation that unknown words or phrases are present in the source document, or that the known translations of source words are based on a very small number of occurrences in the training data. In other cases, translations may be known for individual words in the source document, but not for longer phrases. Translation model adaptation seeks to generate new phrasal translation rules for these source words and phrases. The bias text for a source document may, if comparable, contain a number of English words and phrases that are the English side of these desired rules.

Because the source data and the bias text are not translations of each other and are not sentence aligned, conventional alignment tools, such as GIZA++ (Och and Ney, 2000), cannot be used to align the source and bias text. Because the passages in the bias text are not translations of the source document, it will always be the case that portions of the source document have no translation in the bias text, and portions of the bias text have no translation in the source document. In addition a phrase in one of these texts might have multiple, differing translations in the other text.

Unlike language model adaptation, the entirety of the bias text is not used for translation adaptation. We extract those phrases that occur in at least M of the passages in the bias texts. A phrase is only counted once for every passage in which it occurs, so that repeated use of a phrase within a passage does not affect whether it used to generate new rules. Typically, passages selected by the CLIR tend to be very similar to each other if they are comparable to the source document and are very different from each other if they are not comparable to the source document. Phrases that are identical across passages are the ones that are most likely to be comparable, whereas a phrase or word that occurs in only one passage is likely to be present only by chance or if the passage it is in is not comparable. Filtering the target phrases to those that occur in multiple passages therefore serves not only to reduce the total number of rules, but also to filter out phrases from passages that are not comparable.

For each phrase in the source document we generate a new translation to each of the phrases selected from the bias text, and assign it a low uniform probability.[6] For each translation rule we also have a lexical translation probability that we estimate correctly from the trained word model. These new rules are then added to the phrase table of the existing translation model when translating the source document. Rather than adding probability to the existing generic rules, the new rules are marked as bias rules by the system and given their own feature weight. While the vast majority of these rules are incorrect translations, these incorrect rules will be naturally biased against by the translation system. If the source side of a translation already has a number of observed translations, then the low probability of the new bias rule will cause it to not be selected by the translation system. If the new translation rules would produce garbled English, then it will be biased against by the language model. When this is combined with the language model adapta-

[6] A probability of 1/700 is arbitrarily used for the bias rules although it is then weighted by the bias translation rule weight.

tion, a natural pressure is exerted to use the bias rules for source phrases primarily when it would cause the output to look more like the bias text.

4 Experimental Results

We evaluated the performance of language and translation model adaptation with our translation system on two conditions, the details of which are presented in section 4.1. One condition involved a small amount of parallel training, such as one might find when translating a less commonly taught language (LCTL). The other condition involved the full amount of training available for Arabic-to-English translation. In the case of LCTLs we expect our translation model to have the most deficiencies and be most in need of additional translation rules. So, it is under such a condition we would expect the translation model adaptation to be the most beneficial. We evaluate the system's performance under this condition in section 4.2. The effectiveness of this technique on state-of-the-art systems, and its efficiency when used with a well trained generic translation model is presented in section 4.3.

4.1 Implementation Details

Both language-model and translation-model adaptation are implemented on top of a hierarchical Arabic-to-English translation system with string-to-dependency rules as described in Shen et al. (2008). While generalized rules are generated from the parallel data, rules generated by the translation model adaptation are not generalized and are used only as phrasal rules. A trigram language model was used during decoding, and a 5-gram language model was used to re-score the n-best list after decoding. In addition to the features described in Shen et al. (2008), a new feature is added to the model for the bias rule weight, allowing the translation system to effectively tune the probability of the rules added by translation model adaptation in order to improve performance on the tuning set.

Bias texts were selected from three monolingual corpora: the English Gigaword corpus (2,793,350,201 words), the FBIS corpus (28,465,936 words), and a collection of news archive data collected from the websites of various online, public news sites (828,435,409 words). All three corpora were also part of the generic language model training data. Language model adaptation on both the trigram and 5-gram language models used 10 comparable passages with an interpolation weight of 0.1. Translation model adaptation used 10 comparable passages for the bias text and a value of 2 for M.

Each selected passage contains approximately 300 words, so in the case where 10 comparable passages are used to create a bias text, the resulting text will be 3000 words long on average. The language models created using these bias texts are very specific giving large probability to n-gram sequences seen in those texts.

The construction of the bias texts increases the overall run-time of the translation system, although in practice this is a small expenditure. The most intensive portion is the initial indexing of the monolingual corpus, but this is only required once and can be reused for any subsequent test set that is evaluated. This index can then be quickly searched for comparable passages. When considering research environments, test sets are used repeatedly and bias texts only need to be built once per set, making the building cost negligible. Otherwise, the time required to build the bias text is still small compared to the actual translation time.

All conditions were optimized using BLEU (Papineni et al., 2002) and evaluated using both BLEU and Translation Edit Rate (TER) (Snover et al., 2006). BLEU is an accuracy measure, so higher values indicate better performance, while TER is an error metric, so lower values indicate better performance. Optimization was performed on a tuning set of newswire data, comprised of portions of MTEval 2004, MTEval 2005, and GALE 2007 newswire development data, a total of 48921 words of English in 1385 segments and 173 documents. Results were measured on the NIST MTEval 2006 Arabic Evaluation set, which was 55578 words of English in 1797 segments and 104 documents. Four reference translations were used for scoring each translation.

Parameter optimization method was done using n-best optimization, although the adaptation process is not tied to this method. The MT decoder is run on the tuning set generating an n-best list (where $n = 300$), on which all of the translation features (including bias rule weights) are optimized using

Powell's method. These new weights are then used to decode again, repeating the whole process, using a cumulative n-best list. This continues for several iterations until performance on the tuning set stabilizes. The resulting feature weights are used when decoding the test set. A similar, but simpler, method is used to determine the feature weights after 5-gram rescoring. This n-best optimization method has subtle implications for translation model adaptation. In the first iteration, few bias rules are used in decoding the 300-best, and those that are used frequently help, although the overall gain is small due to the small number of bias rules used. This causes the optimizer to greatly increase the weight of the bias rules, causing the decoder to overuse the bias rules in the next iteration causing a sharp decrease in translation quality. Several iterations are needed for the cumulative n-best to achieve sufficient diversity and size to assign a weight for the bias translation rules that results in an increase in performance over the baseline. Alternative optimization methods could likely circumvent this process. Language model adaptation does not suffer from this phenomenon.

4.2 Less Commonly Taught Language Simulation

In order to better examine the nature of translation model adaptation, we elected to work with a translation model that was trained on only 5 million words of parallel Arabic-English text. Limiting the translation model training in this way simulates the problem of translating less commonly taught languages (LCTL) where less parallel text is available, a situation that is not the case for Arabic. Since the model is trained on less parallel data, it is lacking a large number of translation rules, which is expected to be addressed by the translation model adaptation. By working in an environment with a more deprived baseline translation model, we are giving the translation model adaptation more room to assist.

The experiments described below use a 5 million word Arabic parallel text corpus constructed from the LDC2004T18 and LDC2006E25 corpora. The full monolingual English data were used for the language model and for selection of comparable documents. Unless otherwise specified no language model adaptation was used.

We first establish an upper limit on the gain using translation model adaptation, using the reference data to adapt the translation system. These reference data can be considered to be extremely comparable, better than could ever be hoped to gain by comparable document selection. We first aligned this data using GIZA++ to the source data, simulating the ideal case where we can perfectly determine which source words translate to which comparable words. Because our translation model adaptation system assigns uniform probability to all bias rules, we ignore the correct rule probabilities that we could extract from word alignment and assign uniform probability to all of the bias translation rules. As expected, this gives a large gain over the baseline.

We also examine limiting these new translation rules to those rules whose target side occurs in the top 100 passages selected by CLIR, thus minimizing the adaption to those rules that it theoretically could learn from the bias text. On average, 50% of the rules were removed by this filtering, resulting in a corresponding 50% decrease in the gain over the baseline. The results of these experiments and an unadapted baseline are shown in table 1.

Test Set	TM Adaptation	TER	BLEU
Tune	None	0.4984	0.4080
	Aligned Reference	0.3692	0.5841
	Overlapping Only	0.4179	0.5138
MT06	None	0.5516	0.3468
	Aligned Reference	0.4517	0.5216
	Overlapping Only	0.4899	0.4335

Table 1: LCTL Aligned Reference Adaptation Results

The fair translation model adaptation system, however, does not align source phrases to the correct bias text phrases in such a fashion, and instead aligns all source words to all target words. To investigate the effect of this over production of rules, we again used the reference translations as if they were comparable data, but we ignored the alignments learned by GIZA++, and instead allowed all source phrases to translate to all English phrases in the reference text, with uniform probability. This still shows large gains in translation quality over the baseline, as measured by TER and BLEU. Again, we also examined limiting the text used for translation model adaptation to those phrases that occur in

both the reference text and the top 100 comparable passages selected the CLIR system. While this decreased performance, the system still performs significantly better than the baseline, as shown in the following table 2.

Test Set	TM Adaptation	TER	BLEU
Tune	None	0.4984	0.4080
	Unaligned Ref.	0.4492	0.4566
	Overlapping Only	0.4808	0.4313
MT06	None	0.5516	0.3468
	Unaligned Ref.	0.5254	0.3990
	Overlapping Only	0.5390	0.3695

Table 2: LCTL Unaligned Reference Adaptation Results

Applying translation model and language model adaptation fairly, using only bias text from the comparable data selection, yields smaller gains on both the tuning and MT06 sets, as shown in table 3. The combination of language-model and translation-model adaptation exceeds the gains that would be achieved over the baseline by either method separately.

Test Set	Adaptation	TER	BLEU
Tune	None	0.4984	0.4080
	LM	0.4922	0.4140
	TM	0.4916	0.4169
	LM & TM	0.4888	0.4244
MT06	None	0.5516	0.3468
	LM	0.5559	0.3490
	TM	0.5545	0.3478
	LM & TM	0.5509	0.3536

Table 3: LCTL Fair Adaptation Results

4.3 Full Parallel Training Results

While the simulation described in section 4.2 used only 5 million words of parallel training, 230 million words of parallel data from 18.5 million segments were used for training the full Arabic-to-English translation system. This parallel data includes the LDC2007T08 "ISI Arabic-English Automatically Extracted Parallel Text" corpus (Munteanu and Marcu, 2007), which was created from monolingual corpora in English and Arabic using the algorithm described in Munteanu and Marcu (2005), as the techniques used in that work are separate and independent from the adaptation methods we describe in this paper.[7] Language model adaptation and translation model adaptation were applied both independently and jointly to the translation system, and the results were evaluated against an unadapted baseline, as shown in table 4.

While gains from language model adaptation were substantial on the tuning set, on the MT06 test set they are reduced to a 0.65% gain on BLEU and a negligible improvement in TER. The translation model adaptation performs better with 1.37% improvement in BLEU and a 0.26% improvement in TER. This gain increases to a 2.07% improvement in BLEU and a 0.64% improvement in TER when language adaptation is used in conjunction with the translation model adaptation, showing the importance of using both adaptation methods. While it could be expected that a more heavily trained translation model might not require the benefit of language and translation model adaptation, a more substantial gain over the baseline can be seen when both forms of adaptation are used than in the case with less parallel training—a difference of 2.07% BLEU versus 0.68% BLEU.

Test Set	Adaptation	TER	BLEU
Tune	None	0.4339	0.4661
	LM	0.4227	0.4857
	TM	0.4351	0.4657
	LM & TM	0.4245	0.4882
MT06	None	0.5146	0.3852
	LM	0.5140	0.3917
	TM	0.5120	0.3989
	LM & TM	0.5082	0.4059

Table 4: Full Training Adaptation Results

Of the comparable passages selected by the CLIR system for the MT06 test set in the full training experiment, 16.3% were selected from the News

[7] The two methods are not directly comparable, and so we do not make any attempt to do so. Munteanu and Marcu (2005) creates new parallel corpora from two monolingual corpora. This new parallel data is generally applicable for training a translation model but does not target any particular test set. Our adaptation method does not generate new parallel data, but creates a new specific translation model for a test document that is being translated.

Archive corpus, 81.2% were selected from the English GigaWord corpus and 2.5% were selected from the FBIS corpus. A slightly different distribution was found for the Tuning set, where 17.8% of the passages were selected from the News Archive corpus, 77.1% were selected from the English GigaWord corpus, and 5.1% were selected from the FBIS corpus.

5 Discussion

The reuse of a monolingual corpus that was already used by a translation system for language model training to perform both language and translation model adaptation shows large gains over an unadapted baseline. By leveraging off of a CLIR system, which itself contains no information not already given to the translation system,[8] potentially comparable passages can be found which allow improved translation. Surprisingly, these gains are largest when the baseline model is better trained, indicating that a strong reliance of the adaptation on the existing models.

One explantation for these counter-intuitive results–larger gains in the full training scenario versus the LCTL scenario–is that the lexical probabilities are better estimated in the former case. The bias rules all have equal translation probability and only vary in probability according to the lexical probability of the rules. Better estimates of these lexical probabilities may enable the translation system to more clearly distinguish between helpful and harmful bias rules.

There are many clear directions for the improvement of these methods. The current adaptation method does not utilize the probabilities from the CLIR system and treats the top-ranked passages all as equally comparable regardless of the probability assigned. Variable weighting of passages could prove beneficial to both language model adaptation, where the passages could be weighted proportionally to the probability of the passage being relevant, and translation model adaptation, where the requirement on repetition of phrases across passages could be weighted, as could the probability of the new rules produced by the translation system. In addition, the CLIR score, among other possible features such as phrase overlap, could be used to determine those documents where no comparable passage could be detected and where it would be beneficial to not adapt the models.

A clear limitation of using comparable documents to adapt the language and translation model is that comparable documents must be found. For many source documents, none of the top passages found by the CLIR system were comparable. We suspect that while this will always occur to some extent, this becomes more common as the monolingual data becomes less like the source data, such as when there is a large time gap between the two. The full extent of this and the effect of the level of document comparability on translation remains an open question. In addition, while newswire is an excellent source of comparable text, it is unclear how well this method can be used on newsgroups or spoken data, where the fluency of the source text is diminished. When translating news stories, this technique is not limited to major news events. While many of the events discussed in the source data receive world-wide attention, many are local events that are unreported in the English comparable data used in our experiments. Events of a similar nature or events involving many of the same people often do occur in the English comparable data, allowing improvement even when the stories are quite different.

The adaptation methods described in this paper are not limited to a particular framework of statistical machine translation, but have applicability to any statistical machine translation system that uses a language model or translation rules.

[8] The probabilistic parameters of the CLIR system are estimated from the same parallel corpora that is used to train the generic translation model.

Acknowledgments

This work was supported, in part, by BBN Technologies under the GALE Program, DARPA/IPTO Contract No. HR0011-06-C-0022. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor. We are very grateful to all three reviewers for their careful and thoughtful reviews.

References

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of ACL*, pages 263–270.

Pascale Fung and Lo Yuen Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of COLING-ACL98*, pages 414–420, August.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *Proceedings of EAMT 2005*, Budapest, Hungary, May.

Woosung Kim and Sanjeev Khudanpur. 2003. Cross-Lingual Lexical Triggers in Statistical Language Modeling. In *2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 17–24, July.

Woosung Kim. 2005. *Language Model Adaptation for Automatic Speech Recognition and Statistical Machine Translation*. Ph.D. thesis, The Johns Hopkins University, Baltimore, MD.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based cross-language retrieval. *Information Processing and Management*, 41:523–547.

Yajuan Lu, Jin Huang, and Qun Liu. 2007. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350.

T. Leek Miller and Richard Schwartz. 1998. BBN at TREC7: Using Hidden Markov Models for Information Retrieval. In *TREC 1998*, pages 80–89, Gaithersburg, MD.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31:477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2007. Isi arabic-english automatically extracted parallel text. Linguistic Data Consortium, Philadelphia.

Douglas W. Oard and Bonnie J. Dorr. 1998. Evaluating Cross-Language Text Retrieval Effectiveness. In Gregory Grefenstette, editor, *Cross-Language Information Retrieval*, pages 151–161. Kluwer Academic Publishers, Boston, MA.

F. J. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics*, pages 440–447, Hongkong, China.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Traslation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526.

Philip Resnik and Noah Smith. 2003. The Web as a Parallel Corpus. *Computational Linguistics*, 29:349–380.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, June.

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

Nicola Ueffing. 2006. Using Monolingual Source-Language to Improve MT Performance. In *Proceedings of IWSLT 2006*.

Jinxi Xu, Ralpha Weischedel, and Chanh Nguyen. 2001. Evaluating a Probabilistic Model for Cross-lingual Information Retrieval. In *Proceedings of SIGIR 2001 Conference*, pages 105–110.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language Model Adaptation for Statistical Machine Translation via Structured Query Models. In *Proceedings of Coling 2004*, pages 411–417, Geneva, Switzerland, Aug 23–Aug 27. COLING.