NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Using Monolingual Source-Language Data to Improve MT Performance.

Ueffing, Nicola

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

NRC Publications Record / Notice d'Archives des publications de CNRC:
https://nrc-publications.canada.ca/eng/view/object/?id=de5c3ff8-2697-49bf-8470-347d38d6eee8
https://publications-cnrc.canada.ca/fra/voir/objet/?id=de5c3ff8-2697-49bf-8470-347d38d6eee8

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC·CNRC

Using Monolingual Source-Language Data to Improve MT Performance *

Ueffing, N.
November 2006

* Proceedings of IWSLT 2006. Kyoto, Japan. November 27-28, 2006. NRC 48808.

Copyright 2006 by
National Research Council of Canada

Permission is granted to quote short excerpts and to reproduce figures and tables from this report, provided that the source of such material is fully acknowledged.

Canada

Using Monolingual Source-Language Data to Improve MT Performance

Nicola Ueffing

Interactive Language Technologies Group
National Research Council Canada
Gatineau, Québec, Canada
nicola.ueffing@nrc.gc.ca

Abstract

Statistical machine translation systems are usually trained on large amounts of bilingual text and of monolingual text in the target language. In this paper, we will present a self-training approach which additionally explores the use of monolingual source text, namely the documents to be translated, to improve the system performance. An initial version of the translation system is used to translate the source text. Among the generated translations, target sentences of low quality are automatically identified and discarded. The reliable translations together with their sources are then used as a new bilingual corpus for training an additional phrase translation model. Thus, the translation system can be adapted to the new source data even if no bilingual data in this domain is available. Experimental evaluation was performed on a standard Chinese–English translation task. We focus on settings where the domain and/or the style of the test data is different from that of the training material. We will show a significant improvement in translation quality through the use of the adaptive phrase translation model. BLEU score rises up to 1.1 points, and mWER is reduced by up to 3.1% absolute.

1. Introduction

This paper describes a method for improving an existing statistical machine translation (SMT) system using monolingual source language information. Existing statistical machine translation systems presently benefit from the availability of bilingual parallel or comparable corpora in the source and target language, and from monolingual corpora in the target language. But they do not benefit from the availability of monolingual corpora in the source language. We will show how such corpora can be used to improve the translation performance of the system.

In SMT, the translation process is regarded as a decision problem. Let s_1^J represent a sentence in the source language (the language from which it is desired to translate) and t_1^I represent its translation in the target language. Applying Bayes's Theorem, an SMT system seeks to find a target-language sentence $\hat{t}_1^{\hat{I}}$ that satisfies

$$\arg\max_{t_1^I} p(t_1^I \mid s_1^J) = \arg\max_{t_1^I} p(s_1^J \mid t_1^I) \cdot p(t_1^I) \;, \qquad (1)$$

where $p(t_1^I)$ is the language model, a statistical estimate of the probability of a given sequence of words in the target language. The parameters of the language model are estimated from large text corpora in the target language. The parameters of the target-to-source translation model, $p(s_1^J \mid t_1^I)$, are estimated from a parallel bilingual corpus, in which each sentence expressed in the source language is aligned with its translation in the target language.

State-of-the-art SMT systems basically function as described above, although often other sources of information are combined with the information from $p(s_1^J \mid t_1^I)$ and $p(t_1^I)$ in a log-linear manner. This means that instead of finding a $\hat{t}_1^{\hat{I}}$ that maximizes $p(s_1^J \mid t_1^I) \cdot p(t_1^I)$, these systems search for a $\hat{t}_1^{\hat{I}}$ that maximizes a function of the form

$$p(s_1^J|t_1^I)^{\alpha_0} \cdot p(t_1^I)^{\beta_0} \cdot \prod_{k=1}^{K} g_k^{\alpha_k}(s_1^J, t_1^I) \cdot \prod_{l=1}^{L} h_l^{\beta_l}(t_1^I) \;, \qquad (2)$$

where the feature functions $g_k(s_1^J, t_1^I)$ generate a score based on both source sentence s_1^J and each target hypothesis t_1^I, and feature functions $h_l(t_1^I)$ assess the quality of each t_1^I based on monolingual target-language information. The parameters for functions $g_k(s_1^J, t_1^I)$ can be estimated from bilingual parallel corpora or set by a human designer; the functions $h_l(t_1^I)$ can be estimated from target-language corpora or set by a human designer (and of course, a mixture of all these strategies is possible).

Thus, we see that today's SMT systems benefit from the availability of bilingual parallel corpora for the two relevant languages, since such corpora may be useful in estimating the parameters of the $p(s_1^J \mid t_1^I)$ component and also, possibly, some other bilingual components $g_k(s_1^J, t_1^I)$. Such SMT systems also benefit from the availability of text corpora in the target language, for estimating the parameters of the language model $p(t_1^I)$ and possibly other monolingual target-language components $h_l(t_1^I)$.

However, acquiring monolingual text corpora in the source language is not presently useful in improving an SMT system. In this paper, we present a self-training method which uses monolingual source-language data to improve the performance of an SMT system. It consists of the following steps:

1. Translate new source text using existing MT system,
2. estimate confidence of resulting translations,

3. identify reliable translations based on confidence scores,

4. train new model $g_k(s_1^J, t_1^I)$ on reliable translations and use this as additional feature function in the existing system. In the work presented here, we train a new phrase table on these data.

Through this, the system is provided the ability to adapt to source-language text of a new type (e.g., text discussing new topics not present in the data originally used to train the system, or employing a different style, etc.) without requiring parallel training or development data in the target language. We will show later that translation quality is improved. We will especially consider settings where the test data differ from the training data in domain and/or style.

The paper is structured as follows: Section 2 will describe the proposed method, including details about the baseline SMT system, the identification and filtering of reliable translations, and the training of new phrase tables. Section 3 will explain the experimental setup and present experimental results on a standard Chinese–English translation task. Section 4 will discuss the results and give an outlook of future work.

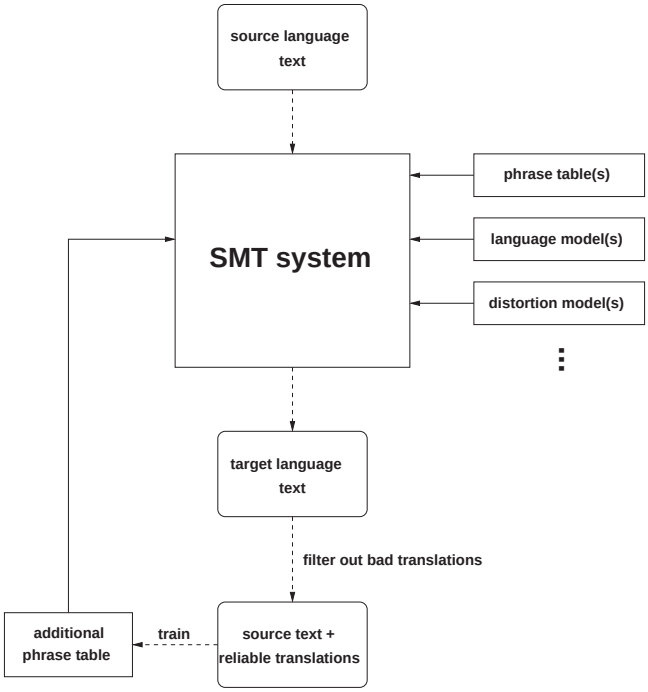

Figure 1: *Schematic diagram of the proposed method.*

2. Use of Monolingual Source Data

2.1. Overview and Motivation

In the following, we will describe the different steps of the proposed self-training method in detail. Figure 1 gives an overview of the process: Some source language text (top) is translated by an existing SMT system, using the different available knowledge sources (right). In the work presented here, we make use of a state-of-the-art phrase-based SMT system. Among the generated translations, the bad ones are automatically determined and removed. The surviving reliable translations, together with their source sentences, are used as new bilingual text to train a phrase table (left). This table is then used by the SMT system as an additional knowledge source.

In the experiments presented in this paper, the method is applied to the development or test corpus to be translated. However, the same can be done if other monolingual source text becomes available. For example, if the SMT system is going to be used for translating newswire text, and a large collection of such data in the source language only (and not in the target language) becomes available, this method can be used to create an additional training corpus. In order to adapt to some test corpus, the relevant parts of the new source data could be identified first, e.g. using information retrieval methods.

The approach of retraining the SMT system on its own translations of a test corpus is a method for adapting the system to this test corpus. It reinforces those phrases in the existing phrase tables which are relevant for translating the new data. Since bad machine translations are filtered out, presumably only phrases of high quality are reinforced whereas the probabilities of low-quality phrase pairs, such as noise in the table or overly confident singletons, degrade. The probability distribution over the phrase pairs should thus get more focused on the (reliable) parts which are relevant for a given test corpus. The method thus provides a means for adapting the existing system to a new domain or style for which no bilingual training or development data is available.

An additional effect of the self-training is that the system can learn new phrase pairs. Suppose that the source phrases 'A B' and 'C D E' each occurred in the parallel corpus used to train the original phrase table $p(s_1^J \mid t_1^I)$, but not contiguously. If in the additional monolingual source-language data the sequence 'A B C D E' occurs frequently, the system has the opportunity to generate new source phrases: e.g., 'A B C', 'B C D E', 'A B C D E'. If the target-language translations generated for these phrases are considered reliable, this enables new bilingual phrases to be learned and put into the new phrase table.

However, the approach has its limitations. It does not enable the system to learn translations of unknown source-language words occurring in the new data. Only words which are already contained in the phrase tables (or equivalent knowledge sources for other types of MT systems) will occur in the newly created bilingual corpus. Furthermore, the approach is limited to the learning of compositional phrases. It is impossible that the system will learn how to translate idioms such as "it is raining cats and dogs" properly into another language, even if correct translations of "it is raining" and "cats and dogs" are contained in the phrase table.

2.2. Baseline MT System

The SMT system which we applied in our experiments is PORTAGE, which is a state-of-the-art phrase-based system.

We will give a short overview in the following; for a detailed description, see [1].

The models (or feature functions) which are employed by the decoder are

- one or several phrase table(s), which model the translation direction $p(s_1^J \mid t_1^I)$,
- one or several n-gram language model(s), trained with the SRILM toolkit [2]. In the experiments reported here, we used 4-gram models,
- a distortion model, which assigns a penalty based on the number of source words which are skipped when generating a new target phrase,
- and a word penalty.

These different models are combined log-linearly as shown in eq. 2. The weights $\alpha_0, \ldots, \alpha_K, \beta_0, \ldots, \beta_L$ are optimized w.r.t. BLEU score [3] using the algorithm described in [4]. This is done on a development corpus which we will call dev1 in the following. The search algorithm implemented in the decoder is a dynamic-programming beam-search algorithm.

After the main decoding step, rescoring with additional models is performed. The system generates a 5,000-best list of alternative translations for each source sentence. These lists are rescored with the following models:

- the different models used in the decoder which are described above,
- two different features based on IBM model 1 [5]: a model-1 probability calculated over the whole sentence, and a feature estimating the number of source words which have a reliable translation. Both features are determined for both translation directions,
- posterior probabilities for words, phrases, n-grams, and sentence length [6, 7]. All of them are calculated over the N-best list and make use of the sentence probabilities which the baseline system assigns to the translation hypotheses.

The weights of these additional models and the models of the decoder are again optimized to maximize BLEU score. This is performed on a second development corpus, dev2.

2.3. Translation of Monolingual Data

Now assume that some new source-language text has become available. In the experiments presented in this paper, these are sentences for which we require translation, i.e. the development or test corpus. Using the baseline SMT system described above, we translate all sentences. For each source sentence, a 5,000-best list of alternative translations is generated by the SMT system. These are used for calculating the confidence scores of the single-best translations as described in the next subsection.

2.4. Confidence Estimation and Filtering

For each single-best translation t_1^I of a new source sentence s_1^J which has been generated by the system, we calculate a confidence score $c(t_1^I)$. Based on this confidence score, it is decided whether t_1^I is a trustworthy translation of s_1^J. To this end, the confidence score $c(t_1^I)$ is compared to a numerical threshold τ. If $c(t_1^I)$ exceeds τ, then the system retains t_1^I as a reliable translation for source sentence s_1^J. If this confidence $c(t_1^I)$ is too low, the translation t_1^I is discarded, and the source sentence s_1^J is left untranslated and not used in the following steps. The threshold τ has been optimized beforehand on a development set. On this set, each translation hypothesis has been labeled as 'correct' or 'incorrect' based on their word error rate (WER). All translations with a WER up to a certain value are considered correct, and all others as incorrect. This follows the method proposed in [8]. These true classes of the machine translations are then used to determine the optimal threshold τ w.r.t. classification error rate.

The confidence $c(t_1^I)$ of the translation t_1^I is computed in the following way. The SMT system generates an N-best list of translation hypotheses for each source sentence. It then estimates the confidence of the single-best translation based on a log-linear combination of the following features:

- a posterior probability which is based on the Levenshtein alignment of the single-best hypothesis over the N-best list [6],
- a posterior probability based on the phrase alignment determined by the SMT system, similar to the source-based posterior probabilities described in [8],
- a language model score determined using an n-gram model.

The log-linear combination of these features is optimized w.r.t. classification error rate on the development set dev1.

However, many other approaches are also possible for calculation of the confidence score, such as those based on other variants of posterior probabilities, more complex translation and language models, methods exploring semantic and syntactic information, etc. See [6, 7, 8, 9] for examples.

Although in the experiments reported here, at most one translation t_1^I for each source sentence s_1^J is retained, it is possible to retain more than one. The same technique can be applied to all hypotheses in the N-best list (or a word graph), allowing for different translations of the same source sentence.

2.5. Phrase Table Training

The new bilingual corpus consisting of the reliable translations and their sources is used to generate a new phrase table for estimating $p(s_1^J \mid t_1^I)$. This is used as a feature function in our decoder. Additionally, we learn the phrase translation model in the opposite translation direction, $p(t_1^I \mid s_1^J)$, which is used for pruning the phrase tables. These two models provide a means of adapting the SMT system to the topic and the style of the new source data.

In our current system, the phrase table training involves first using IBM models [5] for word alignment, and then

Table 1: *Chinese–English Corpora*

corpus	use	# sentences	domains
non-UN	phrase table + LM	3,164,180	news, magazines, laws, Hansards
UN	phrase table + LM	4,979,345	UN Bulletin
English Gigaword	LM	11,681,852	news
multi-p3	dev1	935	news
multi-p4	dev2	919	news
eval-04	test	1,788	newswire (NW), editorials (ED), political speeches (SP)
eval-06	test	3,940	broadcast conversations (BC), broadcast news (BN), newsgroups (NG), newswire (NW)

using the so-called "diag-and" method [10] to extract the phrase pairs that comprise the new phrase table. The maximal length of the phrases was set to 4. We also experimented with longer phrases, but this did not yield any improvement in translation quality. On the original phrase tables, the phrase length restrictions are 5 words for the UN data and 8 for the non-UN data. The new phrase table can be seen as an additional knowledge source which helps the system adapt to the domain or the language of the new source data without requiring bilingual corpora.

2.6. New SMT System

The new phrase table is used as a separate component along with the original phrase tables in a log-linear combination, allowing the system to assign an individual weight to the newly added phrase table. After adding the new table, the weights of all different models in the decoder are optimized on the development corpus dev1 as described in section 2.2. For each new source corpus to be translated, we create an adaptive phrase table as described in the previous subsections. This new phrase table is then plugged into the existing SMT system with this optimized weight. So the new SMT system uses the original phrase tables which are independent of the test corpus, and one new phrase table trained on the test corpus. Again, the weights of the different rescoring models are optimized anew on the development corpus dev2.

3. Experimental Results

3.1. Corpora

We ran experiments on a Chinese–English translation task using the corpora distributed for the NIST MT evaluation (www.nist.gov/speech/tests/mt). The training material used is the one which was permitted for the so-called large-data track in the 2006 NIST evaluation. The Chinese texts have been segmented using the LDC segmenter. The corpora are summarized in table 1. We trained one phrase table on the UN corpus, and another on all other parallel corpora. We also used a subset of the English Gigaword corpus to augment the LM training material.

The multiple translation corpora multi-p3 and multi-p4 were used for optimizing the model weights in the decoder and the rescoring model, respectively. Testing was carried out on the 2004 and 2006 evaluation corpora. They both consist of data from several different domains which are listed in table 1. We see that both the 2004 and the 2006 test corpora contain data from domains which are not covered by the training material. Moreover, the training material consists mainly of written text, whereas the testing is carried out on both text and manually transcribed speech data, e.g. broadcast conversations. The latter have characteristics of spontaneous speech, such as hesitations, repetitions, and incomplete sentences. This will allow us to analyze the adaptation capability of the proposed self-training method.

3.2. Experimental Setup

The test corpora comprise data from different domains. Thus, we trained separate phrase tables on each of these genres. That is, we trained one phrase translation model only on the editorials in the eval-04 corpus and used this for adapting the system to this domain, trained another phrase table on the speeches, etc. The weight for this adaptive phrase table was always the same, namely the one trained on the development corpus dev1. This method yielded three adaptive phrase translation models on the eval-04 corpus, and four on eval-06.

To all our phrase tables, we applied smoothing as proposed in [11]. For the original phrase tables, we used relative frequency estimates smoothed with IBM1 lexical probabilities (the so-called "Zens-Ney method"). The adaptive phrase tables are much smaller than the original ones. In accordance with the findings reported in [11], a different kind of smoothing proved best, namely Zens-Ney-IBM1 smoothing together with Kneser-Ney smoothing. In both cases, the (smoothed) phrase tables were combined log-linearly.

3.3. Results

In the following, we present results measuring translation quality using the BLEU-4 score [3], multi-reference word error rate (mWER) and position-independent word error rate (mPER) [12]. All of them were calculated using our own implementations. The automatic translations as well as the references are lowercased. For the 2004 test corpus, there are 4 references per source sentence. The 2006 evaluation data comprises two parts: the so-called GALE part (consisting of 2,276 source sentences) with 1 reference translation per sentence, and the NIST part (consisting of 1,664 source sentences) with 4 references. Therefore, we will present sepa-

rate scores and error rates on the two subsets of the latter corpus. We will give 95%-confidence intervals for the baseline scores and error rates which have been calculated using bootstrap resampling. Note that BLEU assesses quality, whereas mWER and mPER are error rates. Thus, higher BLEU scores and lower mWER/mPER indicate higher translation quality.

Table 2 presents the overall translation quality on the 2004 and 2006 evaluation corpora. The baseline system described in section 2.2 is compared with the new system which uses the adaptive phrase table as additional knowledge source. We see that on all three sets, translation quality is improved. BLEU score consistently increases, and the error rates drop in all cases. In eight out of those nine cases shown in table 2, the gain in translation quality is significant at the 95%-level. Note that the figures presented in table 2 suggest that the translation quality achieved on the GALE section of the 2006 test set is lower than that on the other two corpora. However, a large part of this difference comes simply from the lack of multiple references for the GALE data. When using only a single reference on the other two sets, we obtain BLEU scores and error rates in the same range as those on the GALE data.

Table 3 presents a detailed analysis of the translation quality, separate for the different genres the test sentences come from. This shows how well the system adapts to the different domains covered by the test data. As the training corpus statistics in table 1 show, the baseline system has been trained mainly on newswire data. Some of the domains of the test data, e.g. the broadcast conversations, are quite different from that w.r.t. topic and style. The results presented in table 3 indicate that the systems which use the additional specific phrase tables adapt well to these data: BLEU score increases in all but two cases, and mPER and mWER decrease for all corpora and genres. The reduction in mWER is significant at the 95%-level in all but two cases. The largest gains are achieved on broadcast conversations and the eval-04 newswire data. Since there is no training/test data mismatch in the latter case, we assume that this is due to an adaptation to the topic. On the eval-04 speeches, there is no gain in translation quality. Interestingly, this is the part of the 2004 evaluation corpus on which the baseline system performs best. Following this thought, we performed an analysis of the improvement in translation quality that the proposed method yields versus the performance of the baseline system on the respective genre which will be shown later in figure 2. Table 3 shows that there is one sub-corpus on which the BLEU score drops slightly, namely the NIST newsgroup data from the 2006 evaluation. mWER, however, is significantly reduced. So we cannot draw any clear conclusions in this case.

Figure 2 plots the relative gain in translation quality (i.e. relative increase in BLEU score or relative decrease in error rate) against the "relative" performance of the baseline system on each genre. The latter is calculated as follows: the difference in BLEU score/mWER on the part of the corpus is divided by the overall BLEU score/mWER on the whole corpus. This gives us a normalized performance measure w.r.t. the full corpus comprising several different genres.

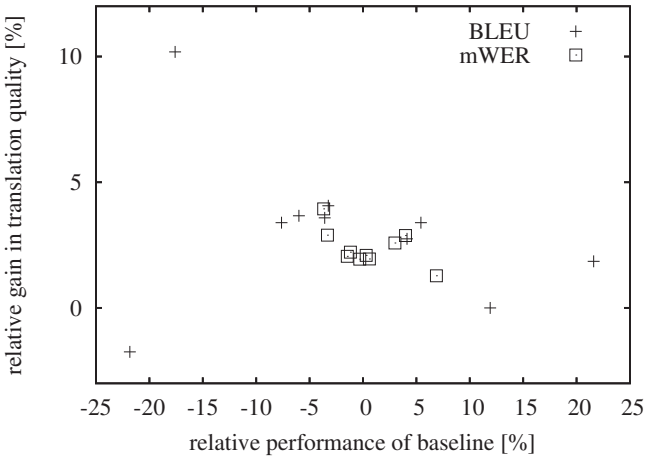

Figure 2: *Relative improvement vs. relative translation quality. One data point per sub-corpus (genre).*

Negative values on the x-axis indicate that for this genre, the translation quality is lower than the overall performance on the whole corpus, and positive values indicate higher translation quality on this genre. For mWER and BLEU score, the plot shows that indeed the lower the relative translation quality on a genre, the higher the gain achieved through adaptation is. The correlation for the mWER data points is 0.55 [1]. For BLEU, however, there is one outlier which refers to the newsgroup data of the 2006 NIST corpus. Removing this outlier yields a correlation of 0.79 (as opposed to 0.25 if all points are included). For mPER, the picture is less clear, with a correlation of 0.36. We therefore did not include mPER in the plot.

In order to see how useful the new phrases are for translation, we analyzed the adaptive phrase tables and the phrases which the SMT system actually used. The statistics are presented in table 4. It shows how many of the machine translations of the new source sentences were considered reliable: In most cases, this is roughly a quarter of the translations. The exception is the broadcast conversation part of the 2006 data where almost half the translations are kept. On these sentence pairs, between 1,900 and 4,000 phrase pairs were learned for the different sub-corpora. The average phrase length is slightly above 2 words for both source and target phrases for all phrase tables (as opposed to an average length of 3-3.5 words in the original phrase tables). To see how useful this new phrase table actually is, we analyzed how many of the phrases which have been learned from the test corpus are used later in generating the best translations (after rescoring). The fourth column shows that for all corpora, about 40% of the phrase pairs from the adaptive model are actually used in translation.

Out of the phrase pairs in the adaptive phrase table, 28% to 48% are entries which are not contained in the original phrase tables. So the system has actually learned new phrases

[1] Note that this value should be taken with a grain of salt because we have only 11 data points.

Table 2: *Translation quality in terms of BLEU score, mWER and mPER on the NIST Chinese–English task.*

corpus		system	BLEU[%]	mWER[%]	mPER[%]
eval-04		baseline	31.8 ±0.7	66.8 ±0.7	41.5 ±0.5
		adapted	32.6	65.3	40.8
eval-06	GALE	baseline	12.7 ±0.5	75.8 ±0.6	54.6 ±0.6
		adapted	13.3	73.6	53.4
	NIST	baseline	27.9 ±0.7	67.2 ±0.6	44.0 ±0.5
		adapted	28.4	65.9	43.4

Table 3: *Translation quality in terms of BLEU score, mWER and mPER. Separate evaluation for each genre on the two test corpora.*

corpus			system	BLEU[%]	mWER[%]	mPER[%]
eval-04		ED	baseline	30.7 ±1.3	67.0 ±1.1	42.3 ±1.0
			adapted	31.8	65.7	41.8
		NW	baseline	30.0 ±1.0	69.1 ±0.9	42.7 ±0.8
			adapted	31.1	67.1	41.8
		SP	baseline	36.1 ±1.4	62.5 ±1.2	38.6 ±0.9
			adapted	36.1	61.7	38.3
eval-06	GALE	BC	baseline	10.8 ±0.7	78.7 ±1.2	59.2 ±1.1
			adapted	11.9	75.6	56.9
		BN	baseline	12.3 ±0.9	76.7 ±1.1	54.0 ±1.1
			adapted	12.8	75.0	53.2
		NG	baseline	11.8 ±1.0	73.6 ±1.0	55.1 ±1.2
			adapted	12.2	71.7	54.3
		NW	baseline	16.2 ±1.1	72.9 ±1.4	49.0 ±1.3
			adapted	16.5	70.8	48.0
eval-06	NIST	BN	baseline	29.5 ±1.4	66.8 ±1.4	44.3 ±1.2
			adapted	30.5	65.5	42.7
		NG	baseline	22.9 ±1.6	68.2 ±1.3	48.8 ±1.1
			adapted	22.5	66.8	48.2
		NW	baseline	29.1 ±1.1	67.0 ±0.8	41.6 ±0.8
			adapted	29.9	65.6	41.4

through self-training. However, an analysis of the number of new phrase pairs which are actually used in translation (presented in the last column of table 4) shows that the newly learned phrases are hardly employed. A comparative experiment showed that removing them from the adapted phrase table yields about the same gain in translation quality as the use of the full adapted phrase table. So the reward from self-training seems to come from the reinforcement of the relevant phrases in the existing phrase tables.

Table 5 presents some translation examples of the baseline and the adapted system. The square brackets indicate phrase boundaries. All examples are taken from the GALE portion of the 2006 test corpus. The domains are broadcast news and broadcast conversation. The examples show that the adapted system outperforms the baseline system both in terms of adequacy and fluency. Especially the third example is interesting: An analysis showed that the target phrase "what we advocate" which is used by the baseline system is an overly confident entry in the original phrase table. The adapted system, however, does not use this phrase here. This indicates that the shorter and more reliable phrases have been reinforced in self-training.

4. Discussion and Outlook

We presented a self-training method which explores monolingual source-language data in order to improve an existing machine translation system. The source data is translated using the MT system, then the reliable translations are automatically identified. Together with their sources, these sentences form a new bilingual corpus which is used to train new translation models. This provides a method of adapting the existing MT system to a new domain or style even if no bilingual training or development data from this domain is available.

Self-training has been explored in other areas of NLP, such as parsing [13] and speech recognition (see below). However, it has not been successfully applied to MT yet to the best of our knowledge. [14] describes a co-training approach for MT which follows a similar spirit: An SMT system for a new language pair is bootstrapped from existing SMT systems in a weakly supervised manner.

It would be interesting to see whether unsupervised training approaches applied in speech recognition [15, 16,

Table 4: *Statistics of the phrase tables trained on the different genres of the test corpora.*

corpus		# sentences	# reliable translations	phrase table size	# adapted phrases used	# new phrases	# new phrases used
eval-04	ED	449	101	1,981	707	679	23
	NW	901	187	3,591	1,314	1,359	47
	SP	438	113	2,321	815	657	25
eval-06	BC	979	477	2,155	759	1,058	90
	BN	1,083	274	4,027	1,479	1,645	86
	NG	898	226	2,905	1,077	1,259	88
	NW	980	172	2,804	1,115	1,058	41

Table 5: *Translation examples from the 2006 GALE corpus (punctuation marks tokenized).*

baseline	[the report said] [that the] [united states] [is] [a potential] [problem] [, the] [practice of] [china 's] [foreign policy] [is] [likely to] [weaken us] [influence] [.]
adapted	[the report] [said that] [this is] [a potential] [problem] [in] [the united states] [,] [china] [is] [likely to] [weaken] [the impact of] [american foreign policy] [.]
reference	the report said that this is a potential problem for america . china 's course of action could possibly weaken the influence of american foreign policy .
baseline	[the capitalist] [system] [, because] [it] [is] [immoral] [to] [criticize] [china] [for years] [, capitalism] [, so] [it] [didn't] [have] [a set of] [moral values] [.]
adapted	[capitalism] [has] [a set] [of] [moral values] [,] [because] [china] [has] [denounced] [capitalism] [,] [so it] [does not] [have] [a set] [of moral] [.]
reference	capitalism , its set of morals , because china has criticized capitalism for many years , this set of morals is no longer there .
baseline	[what we advocate] [his] [name]
adapted	[we] [advocate] [him] [.]
reference	we advocate him .
baseline	[the fact] [that this] [is] [.]
adapted	[this] [is] [the point] [.]
reference	that is actually the point .
baseline	["] [we should] [really be] [male] [nominees] [..] [....]
adapted	[he] [should] [be] [nominated] [male] [,] [really] [.]
reference	he should be nominated as the best actor , really .

17] prove successful in MT as well. That is, a system trained on a small amount of data could be used to translate new source data. The system would then be retrained on the reliable translations together with the original training data. This process can be iterated if more data becomes available over time (as it is the case for newswire data, for example).

The approach presented in this paper trains a phrase translation model on the newly created bilingual corpus. Note that other types of models can also be trained on the new bilingual corpus, for instance, a language model, a distortion model, a sentence length model, etc.

As mentioned earlier, it is also possible to modify the proposed method and retain more than one translation of a source sentence. Often, several correct translations of a source sentence exist, so this approach would allow the system to introduce some more variation into the adapted phrase table. In addition to that, we plan to investigate the relation between the confidence threshold used for filtering out bad translations and the translation quality of the resulting adapted system.

5. Acknowledgments

My thanks go to the PORTAGE team at NRC, esp. George Foster and Roland Kuhn, for their support and valuable feedback.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. 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 Defense Advanced Research Projects Agency (DARPA).

6. References

[1] Johnson, J.H., Sadat, F., Foster, G., Kuhn, R., Simard, M., Joanis, E., Larkin, S., "PORTAGE: with Smoothed Phrase Tables and Segment Choice Models", The North American chapter of the Association for Computational Linguistics (NAACL) Workshop on Statistical Machine Translation. New York City, New York, USA. June

2006.

[2] Stolcke, A., “SRILM - an extensible language modeling toolkit”, Proc. 7th International Conference on Spoken Language Processing (ICSLP), Denver, Colorado, September 2002.

[3] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., “BLEU: A method for automatic evaluation of Machine Translation”, Technical Report RC22176, IBM, September 2001.

[4] Och, F., “Minimum error rate training for statistical machine translation”, Proc. 41th Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan, July 2003.

[5] Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R., “The mathematics of Machine Translation: Parameter estimation”, Computational Linguistics, 19(2), June 1993.

[6] Ueffing, N., Macherey, K., and Ney, H., “Confidence Measures for Statistical Machine Translation”, Proc. Machine Translation Summit IX, New Orleans, LO, September 2003.

[7] Zens, R., and Ney, H., “N-gram Posterior Probabilities for Statistical Machine Translation”, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proc. of the Workshop on Statistical Machine Translation, New York, NY, June 2006.

[8] Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N., “Confidence Estimation for Machine Translation”, Final Report of the Summer Workshop, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 2003.

[9] Quirk, C., “Training a Sentence-Level Machine Translation Confidence Metric”, Proc. 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, May 2004.

[10] Koehn, P., Och, F., and Marcu, D., “Statistical phrase-based translation”, Proc. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL), Edmonton, Alberta, Canada, May 2003.

[11] Foster, G., Kuhn, R., and Johnson, J.H., “Phrasetable Smoothing for Statistical Machine Translation”, Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia, July 2006.

[12] Nießen, S., Och, F., Leusch, G., and Ney, H., ”An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research”, Proc. 2nd International Conference on Language Resources and Evaluation (LREC), Athens, Greece, May-June 2000.

[13] McClosky, D., Charniak, E., and Johnson, M., “Reranking and Self-Training for Parser Adaptation”, Proc. COLING-ACL, Sydney, Australia, 2006.

[14] Callison-Burch, C., and Osborne, M., “Bootstrapping Parallel Corpora”, Proc. NAACL workshop ”Building and Using Parallel Texts: Data Driven Machine Translation and Beyond”, Edmonton, Alberta, Canada, May 2003.

[15] Kemp, T., and Waibel, A., ”Unsupervised Training of a Speech Recognizer: Recent Experiments”, Proc. Eurospeech, Budapest, Hungary, Sep. 1999.

[16] Wessel, F., and Ney, H., ”Unsupervised training of acoustic models for large vocabulary continuous speech recognition”, Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), Trento, Italy, Dec. 2001.

[17] Ma, J., Matsoukas, S., Kimball, O., and Schwartz, R.: ”Unsupervised Training on Large Amounts of Broadcast News Data”, Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 3, Toulouse, May 2006.