Investigations on Translation Model Adaptation Using Monolingual Data

Patrik Lambert, Holger Schwenk, Christophe Servan, Sadaf Abdul-Rauf

Investigations on Translation Model Adaptation Using Monolingual Data

Patrik Lambert, Holger Schwenk, Christophe Servan and Sadaf Abdul-Rauf
LIUM, University of Le Mans
72085 Le Mans, France
FirstName.LastName@lium.univ-lemans.fr

Abstract

Most of the freely available parallel data to train the translation model of a statistical machine translation system comes from very specific sources (European parliament, United Nations, etc). Therefore, there is increasing interest in methods to perform an adaptation of the translation model. A popular approach is based on unsupervised training, also called self-enhancing. Both only use monolingual data to adapt the translation model. In this paper we extend the previous work and provide new insight in the existing methods. We report results on the translation between French and English. Improvements of up to 0.5 BLEU were observed with respect to a very competitive baseline trained on more than 280M words of human translated parallel data.

1 Introduction

Adaptation of a statistical machine translation system (SMT) is a topic of increasing interest during the last years. Statistical (n-gram) language models are used in many domains and several approaches to adapt such models were proposed in the literature, for instance in the framework of automatic speech recognition. Many of these approaches were successfully used to adapt the language model of an SMT system. On the other hand, it seems more challenging to adapt the other components of an SMT system, namely the translation and reordering models. In this work we consider the adaptation of the translation model of a phrase-based SMT system.

While rule-based machine translation rely on rules and linguistic resources built for that purpose, SMT systems can be developed without the need of any language-specific expertise and are only based on bilingual sentence-aligned data ("*bitexts*") and large monolingual texts. However, while monolingual data are usually available in large amounts and for a variety of tasks, bilingual texts are a sparse resource for most language pairs.

Current parallel corpora mostly come from one domain (proceedings of the Canadian or European Parliament, or of the United Nations). This is problematic when SMT systems trained on such corpora are used for general translations, as the language jargon heavily used in these corpora is not appropriate for everyday life translations or translations in some other domain. This problem could be attacked by either searching for more in-domain training data, e.g. by exploring comparable corpora or the WEB, or by adapting the translation model to the task. In this work we consider translation model adaptation without using additional *bilingual data*. One can distinguish two types of translation model adaptation: first, adding new source words or/and new translations to the model; and second, modifying the probabilities of the existing model to better fit the topic of the task. These two directions are complementary and could be simultaneously applied. In this work we focus on the second type of adaptation.

In this work, we focus on statistical phrase-based machine translations systems (PBSMT), but the methods could be also applied to hierarchical systems. In PBSMT, the translation model is represented by a large list of all known source phrases and their translations. Each entry is weighted using several probabilities, e.g. the popular Moses

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 284–293,
Edinburgh, Scotland, UK, July 30–31, 2011.

system uses phrase translation probabilities in the forward and backward direction, as well as lexical probabilities in both directions. The entries of the phrase-table are automatically extracted from sentence aligned parallel data and they are usually quite noisy. It is not uncommon to encounter several hundreds, or even thousands of possible translations of frequent source phrases. Many of these automatically extracted translations are probably wrong and are never used since their probabilities are (fortunately) small in comparison to better translations. Therefore, several approaches were proposed to filter these phrase-tables, reducing considerably their size without any loss of the quality, or even achieving improved performance (Johnson et al., 2007).

Given these observations, adaptation of the translation model of PBSMT systems could be performed by modifying the probability distribution of the existing phrases without necessarily modifying the entries. The idea is of course to increase the probabilities of translations that are appropriate to the task and to decrease the probabilities of the other ones. Ideally, we should also add new translations or source phrase, but this seems to be more challenging without any additional parallel data.

A common way to modify a statistical model is to use a mixture model and to optimize the coefficients to the adaptation domain. This was investigated in the framework of SMT by several authors, for instance for word alignment (Civera and Juan, 2007), for language modeling (Zhao et al., 2004; Koehn and Schroeder, 2007) and to a lesser extent for the translation model (Foster and Kuhn, 2007; Chen et al., 2008). This mixture approach has the advantage that only few parameters need to be modified, the mixture coefficients. On the other hand, many translation probabilities are modified at once and it is not possible to selectively modify the probabilities of particular phrases.

Another direction of research is self-enhancing of the translation model. This was first proposed by Ueffing (2006). The idea is to translate the test data, to filter the translations with help of a confidence score and to use the most reliable ones to train an additional small phrase table that is jointly used with the generic phrase table. This could be also seen as a mixture model with the in-domain component being build on-the-fly for each test set. In practice, such an approach is probably only feasible when large amounts of test data are collected and processed at once, e.g. a typical evaluation set up with a test set of about 50k words. This method of self-enhancing the translation model seems to be more difficult to apply for on-line SMT, e.g. a WEB service, since often the translation of some sentences only is requested. In follow up work, this approach was refined (Ueffing et al., 2007). Domain adaptation was also performed simultaneously for the translation, language and reordering model (Chen et al., 2008).

A somehow related approach was named lightly-supervised training (Schwenk, 2008). In that work an SMT system is used to translate large amounts of monolingual texts, to filter them and to add them to the translation model training data. This approach was reported to obtain interesting improvements in the translations quality (Schwenk and Senellart, 2009; Bertoldi and Federico, 2009). In comparison to *self enhancing* as proposed by Ueffing (2006), lightly-supervised training does not adapt itself to the test data, but large amounts of monolingual training data are translated and a completely new model is built. This model can be applied to any test data, including a WEB service.

In this paper we propose to extend this approach in several ways. First, we argue that the automatic translations should not be performed from the source to the target language, but in the opposite direction. Second, we propose to use the segmentation obtained during translation instead of performing word alignments with GIZA++ (Och and Ney, 2003) of the automatic translations. Finally, we propose to enrich the vocabulary of the adapted system by detecting untranslated words and automatically inferring possible translations from the stemmed form and the existing translations in the phrase table.

This paper is organized as follows. In the next section we first describe our approach in detail. Section 3 describes the considered task, the available resources and the baseline PBSMT system. Results are summarized in section 4 and the paper concludes with a discussion and perspectives of this work.

2 Architecture of the approach

In this paper we propose to extend in several ways the translation model adaptation by unsupervised

training as proposed by Schwenk (2008). In that paper the authors propose to first build a PBSMT system using all available human translated bitexts. This system is then used to translate large amounts of monolingual data in the source language. These automatic translations are filtered using the sentence-length normalized log score of Moses, i.e. the sum of the log-scores of all feature functions. Putting a threshold on this score, only the most reliable translations are kept. This threshold was determined experimentally. The automatic translations were added to the parallel training data and a new PBSMT model was build, performing the complete pipeline of word alignment with GIZA++, phrase extraction and scoring and tuning the system on development data with MERT. In Schwenk (2009) significant improvement were obtained by this approach when translating from Arabic to French.

2.1 Choice of the translation direction

First, we argue that it should be better to translate monolingual data in the opposite translation direction of the system that we want to improve, i.e. from the target into the source language. When translating large amounts of monolingual data, the system will of course produce some wrong translations with respect to choice of the vocabulary, to word order, to morphology, etc. If we translate from the source to the target language, these wrong translations are added to the phrase table and may be used in future translations performed by the adapted system. When we add the automatic translations performed in the opposite direction to the training data, the possibly wrong translations will appear on the source side of the entries in the adapted phrase table. PBSMT systems segment the source sentence according to the available entries in the phrase table. Since the source sentence is usually grammatically and semantically correct, with the eventual exception of speech translation, it is unlikely that the wrong entries in the phrase table will be ever used, e.g. phrases with bad word choice or wrong morphology.

The question of the choice of the translation direction was already raised by Bertoldi and Federico (2009). However, when data in the source language is available they adapt only the translation model (TM), while they adapt the TM *and the language model* (LM) when data in the target language is given. Of course the system with adapted LM is much better, but this doesn't prove that target monolingual data are better than source monolingual data for TM adaptation. In our paper, we use the same, best, LM for all systems and we adapt the baseline system with bitexts synthesized from source or target monolingual data.

2.2 Word alignment

In the work of Schwenk (2008), the filtered automatic translation were added to the parallel training data and the full pipeline to build a PBSMT system was performed again, including word alignment with GIZA++. Word alignment of bitexts of several hundreds of millions of words is a very time consuming step. Therefore we propose to use the segmentation into phrases and words obtained implicitly during the translation of the monolingual data with the moses toolkit. These alignments are simply added to the previously calculated alignments of the human translated bitexts and a new phrase table is built.

This new procedure does not only speed-up the overall processing, but there are also investigations that these alignments obtained by decoding are more suitable to extract phrases than the symmetrized word alignments produced by GIZA++. For instance, Wuebker et al. (2010) proposed to translate the *training data*, using forced alignment and a leave-one-out technique, and to use the induced alignments to extract phrases. They have observed improvements with respect to word alignment obtained by GIZA++. On the other hand, Bertoldi and Federico (2009) adapted an SMT system with automatic translations and trained the translation and reordering models on the word alignment used by moses. They reported a very small drop in performance with respect to training word alignments with GIZA++. Similar ideas were also used in pivot translation. Bertoldi et al. (2008) translated from the pivot language to the source language to create parallel training data for the direct translation.

2.3 Treatment of unknown words

Statistical machine translation systems have some trouble dealing with morphologically rich languages. It can happen, in function of the available training data, that translations of words are only

Source language French	Source language stemmed form	Target language English
finies	fini	finished
effacés	effacé	erased
hawaienne	hawaien	Hawaiian
...	...	...

Table 1: Example of translations from French to English which are automatically extracted from the phrase-table with the stemmed form.

known in some forms and not in others. For instance, for a user of MT technology it is quite difficult to understand why the system can translate the French word "je pense"[1], but not "tu penses"[2]. There have been attempts in the literature to address this problem, for instance by Habash (2008) to deal with the Arabic language. It is actually possible to automatically infer possible translations when translating from a morphologically rich language, to a simpler language. In our case we use this approach to translate from French to English.

Several of the unknown words are actually adjectives, nouns or verbs in a particular form that itself is not known, but the phrase table would contain the translation of a different form. As an example we can mention the French adjective *finies* which is in the female plural form. After stemming we may be able to find the translation in a dictionary which is automatically extracted from the phrase-table (see Table 1). This idea was already outlined by (Bojar and Tamchyna, 2011) to translate from Czech to English.

First, we automatically extract a dictionary from the phrase table. This is done, be detecting all 1-to-1 entries in the phrase table. When there are multiple entries, all are kept with their lexical translations probabilities. Our dictionary has about 680k unique source words with a total of almost 1M translations.

source segment	les travaux sont **finis**
stemmed	les travaux sont **fini**
segment proposed	les travaux sont <n translation="finished||ended" prob="0.008||0.0001">**finis**</n>

Table 2: Example of the treatment of an unknown French word and its automatically inferred translation.

The detection of unknown words is performed by comparing the *n*-grams contained in the phrase table and the source segment in order to detect identical words. Once the unknown word is selected, we are looking for its stemmed form in the dictionary and propose some translations for the unknown word based on lexical score of the phrase table (see Table 2 for some examples). The stemmer used is the snowball stemmer[3]. Then the different hypothesis are evaluated with the target language model.

This kind of processing could be done either before running the Moses decoder, *i.e.* using the XML mark-up of Moses, or after decoding by post-processing the untranslated words. In both cases, we are unable to differentiate the possible translations of the same source phrase with meaningful translation probabilities, and they won't be added to the phrase-table, nor put into a context with other words that may trigger their use.

Therefore, we propose to use this technique to replace unknown words during the translation of the monolingual data that we use to adapt the translation model. By these means, the automatically induced translations of previously unknown morphological forms will be put into a context and actually appear in the new adapted phrase-table. The corresponding translation probabilities will be those corresponding to their frequency in the monolingual in-domain data.

This procedure has been implemented, but we were not able to obtain improvements in the BLEU score. However, one can ask if automatic metrics, evaluated on a test corpus of limited size, are the best choice to judge this technique. In fact, in our setting we have observed that less than 0.2% of the words in the test set are unknown. We argue that the ability to complement the phrase-table with many morphological forms of other wise known words, can only improve the usability of SMT systems.

3 Task Description and resources

In this paper, we consider the translation of news texts between French and English, in both directions. In order to allow comparisons, we used exactly the same data as those allowed for the international evaluation organized in the framework of the sixth workshop on SMT, to be held in Edinburgh

[1] I think

[2] you think

[3] http://snowball.tartarus.org/

Parallel data	Size [M words]	English/French		French/English	
		Dev	Test	Dev	Test
Eparl + nc	54	26.20 (0.06)	28.06 (0.2)	26.70 (0.06)	27.41 (0.2)
Eparl + nc + crawled1	168	26.84 (0.09)	29.08 (0.1)	27.96 (0.09)	28.20 (0.04)
Eparl + nc + crawled2	**286**	**26.95 (0.04)**	**29.29 (0.03)**	**28.20 (0.03)**	**28.57 (0.1)**
Eparl + nc + un	379	26.57	28.52	-	-
Eparl + nc + crawled1 + un	514	26.87	28.99	-	-
Eparl + nc + crawled2 + un	631	26.99	29.26	-	-

Table 4: Case sensitive BLEU scores as a function of the amount of parallel training data. (Eparl=Europarl, nc=News Commentary, crawled1/2=sub-sampled crawled bitexts, un=sub-sampled United Nations bitexts).

Corpus	English	French
Bitexts:		
Europarl	50.5M	54.4M
News Commentary	2.9M	3.3M
United Nations	344M	393M
Crawled (10^9 bitexts)	667M	794M
Development data:		
newstest2009	65k	73k
newstest2010	62k	71k
Monolingual data:		
LDC Gigaword	4.1G	920M
Crawled news	2.6G	612M

Table 3: Available training data for the translation between French and English for the translation evaluation at WMT'11 (number of words after tokenisation).

in July 2011. Preliminary results of this evaluation are available on the Internet.[4] Table 3 summarizes the available training and development data. We optimized our systems on `newstest2009` and used `newstest2010` as internal test set. For both corpora, only one reference translations is available. Scoring was performed with NIST's implementation of the BLEU score ('mt-eval' version 13).

3.1 Baseline system

The baseline system is a standard phrase-based SMT system based on the the Moses SMT toolkit (Koehn et al., 2007). It uses fourteen features functions for translation, namely phrase and lexical translation probabilities in both directions, seven features for the lexicalized distortion model, a word and a phrase penalty, and a target language model. It is constructed as follows. First, word alignments in both directions are calculated. We used a multi-threaded version of the GIZA++ tool (Gao and Vogel, 2008). Phrases and lexical reorderings are extracted using the default settings of the Moses toolkit. All the bitexts were concatenated. The parameters of Moses are tuned on the development data using the MERT tool. For most of the runs, we performed three optimizations using different starting points and report average results. English and French texts were tokenised using a modified version of the tools of the Moses suite. Punctuation and case were preserved.

The language models were trained on all the available data, i.e. the target side of the bitexts, the whole Gigaword corpus and the crawled monolingual data. We build 4-gram back-off LMs with the SRI LM toolkit using Modified Kneser-Ney and no cut-off on all the n-grams. Past experience has shown that keeping all n-grams slightly improves the performance although this produces quite huge models (10G and 30G of disk space for French and English respectively).

Table 4 gives the baseline results using various amounts of bitexts. Starting with the Europarl and the News Commentary corpora, various amounts of human translated data were added. The organizers of the evaluation provide the so called 10^9 French-English parallel corpus which contains almost 800 million words of data crawled from Canadian and European Internet pages. Following works from the 2010 WMT evaluation (Lambert et al., 2010), we filtered this data using IBM-1 probabilities and language model scores to keep only the most reliable translations. Two subsets were built with 115M and 232M English words respectively (using two differ-

[4] http://matrix.statmt.org

alignment	Dev	Test	
	BLEU	BLEU	TER
giza	27.34 (0.01)	29.80 (0.06)	55.34 (0.06)
reused giza	27.40 (0.05)	29.82 (0.10)	55.30 (0.02)
reused moses	27.42 (0.02)	29.77 (0.06)	55.27 (0.03)

Table 5: Results for systems trained via different word alignment configurations. The values are the average over 3 MERT runs performed with different seeds. The numbers in parentheses are the standard deviation of these three values. Translation was performed from English to French, adding 45M words of automatic translations (translated from French to English) to the baseline system "eparl+nc+crawled2".

ent settings of the filter thresholds). They are referred to as "crawled1" and "crawled2" respectively. Adding this data improved the BLEU score of almost 1 BLEU point (28.30 → 29.27). This is our baseline system to be improved by translation model adaptation. Using the UN data gave no significant improvement despite its huge size. This is probably a typical example that it is not necessarily useful to use all available parallel training data, in particular when a very specific (out-of domain) jargon is used. Consequently, the UN data was not used in the subsequent experiments.

We were mainly working on the translation from English to French. Therefore only one baseline system was build for the reverse translation direction.

4 Experimental Evaluation

The system trained on Europarl, News Commentary and the sub-sampled version of the 10^9 bitexts ("eparl+nc+crawled2", in the third line of Table 3), was used to translate parts of the crawled news in French and English. Statistics on the translated data are given in Table 6.

We focused on the most recent data since the time period of our development and test data was end of 2008 and 2009 respectively. In the future we will translate all the available monolingual data and make it available to the community in order to ease the widespread use of this kind of translation model adaptation methods. These automatic translations were filtered using the sentence normalized log-score of the decoder, as proposed by (Schwenk, 2008). However, we did not perform systematic experiments to find the optimal threshold on this score, but simply used a value which seems to be a good compromise of quality and quantity of the translations. This gave us about 45M English words of automatic translations from French, as well as the translations into French of 100M English words, to be used to adapt the baseline systems.

Corpus	French (fe)		English (ef)	
	available	filtered	available	filtered
2009	92	31	121	45
2010	43	12	112	49
2011	8	2	15	6
total	219	45	177	100

Table 6: Monolingual data used to adapt the systems, given in millions of English words. Under "French (fe)", we indicated the number of translated English words from French, and under "English (ef)" we reported the number of source English words translated into French. Thus "fe" and "ef" refer respectively to French–English and English–French translation direction of monolingual data. In the experiments we used the 100M English–French (ef) filtered monolingual data, as well as a 45M-word subset (in order to have the same amount of data as for French–English) and a 65M-word subset.

4.1 Word alignment

In order to build a phrase table with the translated data, we re-used the word alignment obtained during the translation with the moses toolkit. We compared the system trained via these alignments to the systems built by running GIZA++ on all the data. When word alignments of the baseline corpus (not adapted) are trained together with the translated data, they could be affected by phrase pairs coming from incorrect translations. To measure this effect, we trained an additional system, for which the alignments of the baseline corpus are those trained without the translated data. For the translated data, we re-use the GIZA++ alignments trained on all the data. Results for these three alignment configura-

baseline	translated bitexts	Dev	Test	
		BLEU	BLEU	TER
Eparl + nc	-	26.20 (0.06)	28.06 (0.22)	56.85 (0.09)
	news fe 45M	**27.18 (0.09)**	**29.03 (0.07)**	**55.97 (0.07)**
	news ef 45M	26.15 (0.04)	28.44 (0.09)	56.56 (0.11)
Eparl + nc + crawled2	-	26.95 (0.04)	29.29 (0.03)	55.77 (0.19)
	news fe 45M	**27.42 (0.02)**	**29.77 (0.06)**	**55.27 (0.03)**
	news ef 45M	26.75 (0.04)	28.88 (0.10)	56.06 (0.05)

Table 7: Translation results of the English–French systems augmented with a bitext obtained by translating news data from English to French (ef) and French to English (fe). 45M refers to the number of English running words.

baseline	translated bitexts	Dev	Test	
		BLEU	BLEU	TER
Eparl + nc	-	26.70 (0.06)	27.41 (0.24)	55.07 (0.17)
	news fe 45M	27.47 (0.08)	27.77 (0.23)	54.84 (0.13)
	news ef 45M	27.55 (0.05)	28.51 (0.10)	54.12 (0.09)
	news ef 65M	27.58 (0.03)	28.70 (0.09)	54.06 (0.17)
	news ef 100M	**27.63 (0.06)**	**28.68 (0.06)**	**54.02 (0.06)**
Eparl + nc + crawled2	-	28.20 (0.03)	28.54 (0.12)	54.17 (0.15)
	news fe 45M	28.02 (0.11)	28.40 (0.10)	54.45 (0.06)
	news ef 45M	28.24 (0.06)	28.93 (0.22)	53.90 (0.08)
	news ef 65M	28.16 (0.19)	28.75 (0.06)	54.03 (0.14)
	news ef 100M	**28.28 (0.09)**	**28.96 (0.03)**	**53.79 (0.09)**

Table 8: Translation results of the French–English systems augmented with a bitext obtained by translating news data from English to French (ef) and French to English (fe). 45M/65M/100M refers to the number of English running words.

tions are presented in Table 5. In these systems French sources and English translations (45 million words) were added to the "eparl+nc+crawled2" baseline corpus. According to BLEU and TER metrics, reusing Moses alignments to build the adapted phrase table has no significant impact on the system performance. We repeated the experiment without the 10^9 corpus and with the smaller selection of 10^9 (`crawled1`) and arrived to the same conclusion. However, the re-use of Moses alignments saves time and resources. On the larger baseline corpus, the mGiza process lasted 46 hours with two jobs of 4 thread running and a machine with two Intel X5650 quad-core processors.

4.2 Choice of the translation direction

A second point under study in this work is the effect of the translation direction of the monolingual data used to adapt the translation model. Tables 7 and 8 present results for, respectively, English–French and French–English systems adapted with news data translated from English to French (ef) and French to English (fe). The experiment was repeated with two baseline corpora. The results show clearly that target to source translated data are more useful than source to target translated data. The improvement in terms of BLEU score due to the use of target-to-source translated data instead of source-to-target translated data ranges from 0.5 to 0.9 for the French–English and English–French systems. For instance, when translating from English to French (Table 7), the baseline system "eparl+nc" achieves a BLEU score of 28.06 on the test set. This could be improved to 29.03 using automatic translations in the reverse direction (French to English), while we only achieve a BLEU score of 28.44 when using automatic translation performed in the same direction as the system to be adapted. The effect is even clearer when we try to adapt the large system

"eparl+nc+crawled2". Adding automatic translations translated from English-to-French did actually lead to a lower BLEU score (29.29 → 28.88) while we observe an improvement of nearly 0.5 BLEU in the other case.

With target-to-source translated news data, the gain with respect to the baseline corpus for English-French systems (Table 7) is nearly 1 BLEU for "Eparl+nc" and 0.5 BLEU for "Eparl+nc+crawled2". With the same amount of translated data (45 million English words), approximately the same gains are observed in French–English systems. Due to the larger availability of English news data, we were able to use larger sets of target-to-source translated data for French-English systems, as can be seen in Table 8. With a bitext containing additionally 20 million English words, we get a further improvement of 0.2 BLEU for "Eparl+nc" (28.51 → 28.70), but no improvement for "Eparl+nc+crawled2" (the BLEU score is even lower, but the scores lie within the error interval). No further gain on the test data is achieved if we add again 35 million English words (total of 100M words) to the system "Eparl+nc". With the "Eparl+nc+crawle2" baseline, no significant improvement is observed if we adapt the system with 100M words instead of only 45M.

4.3 Result analysis

To get more insight into what happens to the model when we add the automatic translations, we calculated some statistics of the phrase table, presented in Table 9. Namely, we calculated the number of entries in the phrase table, the average number of translation options of each source phrase, the average entropy for each source phrase, the average source phrase length (in words) and the average target phrase length. The entropy is calculated over the probabilities of all translation options for each source phrase. Comparing the baseline with "Eparl+nc" and the baseline with "Eparl+nc+crawl2", we can observe that the average number of translation options was nearly multiplied by 3 with the addition of 230 million words of human translated bitexts. As a consequence the average entropy was increased from 1.84 to 2.08. On the contrary, adding 100 million words of in-domain automatic translations, the average number of translation options increased by only 5% for the "Eparl+nc" baseline, and decreased for the "Eparl+nc+crawl2" baseline. A decrease may occur if new source phrases with less translation options than the average are added. Furthermore, with the addition of 45 million words of in-domain data, the average entropy dropped from 1.84 to 1.33 or 1.60 for the "Eparl+nc" baseline, and from 2.08 to 1.81 or 1.96 for the "Eparl+nc+crawl2" baseline. With both baselines, the more translations are added to the system, the lower the entropy, although in some case the number of translation options increases (this is the case when we pass from 65M to 100M words of synthetic data). These results illustrate the fact that the automatic translations only reinforce some probabilities in the model, with the subsequent decrease in entropy, while human translations add new vocabulary. Note also that in the corpus using automatic translations, new words can only occur in the source side. Thus when translating from French to English, automatic translations from English to French are expected to yield more translation options and a higher entropy than the automatic translations from French to English. This is what is effectively observed in Table 9.

5 Conclusion

Unsupervised training is widely used in other areas, in particular large vocabulary speech recognition. The statistical models in speech recognition use a *generative approach* based on small units, usually triphones. Each triphone is modeled by a hidden Markov model and Gaussian mixture probability distributions (plus many improvements like parameter tying etc). Many methods were developed to adapt such models. The corresponding model in statistical machine translation is the phrase table, a long list of known words with their translations and probabilities. It seems much more challenging to adapt this kind of statistical model with unsupervised training, i.e. monolingual data. Nevertheless, we believe that unsupervised training can be also very useful in SMT. To the best of our knowledge, work in this area is very recent and only in its beginnings. This paper tries to give additional insights in this promising method.

Our work is based on the approach initially pro-

baseline	translated bitexts	entries (M)	translations	entropy	src size	trg size
Eparl + nc	-	7.16	83.83	1.84	1.80	2.81
	news fe 45M	7.42	70.00	1.33	1.83	2.80
	news ef 45M	8.24	81.58	1.60	1.86	2.79
	news ef 65M	8.42	81.58	1.55	1.88	2.79
	news ef 100M	9.21	85.93	1.54	1.90	2.79
Eparl + nc + crawl2	-	25.42	235.16	2.08	1.76	2.93
	news fe 45M	25.54	217.21	1.81	1.77	2.93
	news ef 45M	26.09	228.07	1.96	1.78	2.93
	news ef 65M	26.21	226.45	1.91	1.78	2.93
	news ef 100M	26.79	227.08	1.89	1.79	2.93

Table 9: Phrase table statistics for French–English systems augmented with bitexts built via automatic translations. Only the entries useful to translate the development set were present in the considered phrase table.

posed in (Schwenk, 2008): build a first SMT system, use it to translate large amounts of monolingual data, filter the obtained translations, add them to the bitexts and build a new system from scratch.

We proposed several extensions to this technique which seem to improve the translations quality in our experiments. First of all, we have observed that it is clearly better to add automatically translated texts to the translations model training data which were translated from the target to the source language. This seems to ensure that potentially wrong translations are not used in the new model.

Second, we were able to skip the process of performing word alignment of this additional parallel data without any significant loss in the BLEU score. Performing word alignments with GIZA++ can easily take several days when several hundred millions of bitexts are available. Instead, we directly used the word alignments produced by Moses when translating the monolingual data. This resulted in an appreciable speed-up of the procedure, but has also interesting theoretical aspects. Reusing the word alignment from the translation process is expected to result in a phrase extraction process that is more consistent with the use of the phrases.

Finally, we outlined a method to automatically add new translations without any additional parallel training data. In fact, when translating from a morphologically rich language to an easier one, in our case from French to English, it is often possible to infer the translations of unobserved morphological forms of nouns, verbs or adjectives. This is obtained by looking up the stemmed form in an automatically constructed dictionary. This kind of approach could be also applied to a classical PBSMT system, by adding various forms to the phrase table, but it is not obvious to come up with reasonable translations probabilities for these new entries. In our approach, the unknown word forms are processed in large amounts of monolingual data and the induced translations will appear in the context of complete sentences. Wrong translations can be blocked by the language model and the new translations can appear in phrases of various lengths.

This paper provided a detailed experimental evaluation of these methods. We considered the translation between French and English using the same data than was made available for the 2011 WMT evaluation. Improvement of up to 0.5 BLEU were observed with respect to an already competitive system trained on more than 280M words of human translated parallel data.

Acknowledgments

This work has been partially funded by the French Government under the project COSMAT (ANR ANR-09-CORD-004.) and the European Commission under the project EUROMATRIXPLUS (ICT-2007.2.2-FP7-231720).

References

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation. In *Forth Workshop on SMT*, pages 182–189.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. In *IWSLT*, pages 143–149.

Ondřej Bojar and Aleš Tamchyna. 2011. Forms Wanted: Training SMT on Monolingual Data. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel, January.

Boxing Chen, Min Zhang, Aiti Aw, and Haizhou Li. 2008. Exploiting n-best hypotheses for SMT self-enhancement. In *ACL*, pages 157–160.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Second Workshop on SMT*, pages 177–180, June.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *EMNLP*, pages 128–135.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *ACL 08*.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *EMNLP*, pages 967–975, Prague, Czech Republic.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Second Workshop on SMT*, pages 224–227, June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, demonstration session*.

Patrik Lambert, Sadaf Abdul-Rauf, and Holger Schwenk. 2010. LIUM SMT machine translation system for WMT 2010. In *Workshop on SMT*, pages 121–126.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignement models. *Computational Linguistics*, 29(1):19–51.

Holger Schwenk and Jean Senellart. 2009. Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In *MT Summit*.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *IWSLT*, pages 182–189.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *ACL*, pages 25–32.

Nicola Ueffing. 2006. Using monolingual source-language data to improve MT performance. In *IWSLT*, pages 174–181.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *ACL*, pages 475–484, Uppsala, Sweden, July. Association for Computational Linguistics.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Coling*.