Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation

Holger Schwenk

LIUM, University of Le Mans, FRANCE
Holger.Schwenk@lium.univ-lemans.fr

Abstract

Sentence-aligned bilingual texts are a crucial resource to build statistical machine translation (SMT) systems. In this paper we propose to apply lightly-supervised training to produce additional parallel data. The idea is to translate large amounts of monolingual data (up to 275M words) with an SMT system, and to use those as additional training data. Results are reported for the translation from French into English. We consider two setups: first the intial SMT system is only trained with a very limited amount of human-produced translations, and then the case where we have more than 100 million words. In both conditions, lightly-supervised training achieves significant improvements of the BLEU score.

1. Introduction

Statistical machine translation (SMT) is today considered as a serious alternative to rule-based machine translation (RBMT). While RBMT systems 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 is usually available in large amounts, bilingual texts are a sparse resource for most of the language pairs. The largest SMT systems are currently built for the translation of news material from Mandarin and Arabic into English, using more than 170M words of bitexts that are easily available from the LDC. The possibility to develop a MT system using only aligned bilingual texts is generally mentioned as an advantage of SMT systems. On the other hand, this can also be a handicap for this approach. For some language pairs bilingual corpora just do not exist, e.g. Japanese/Spanish, or the existing corpora are too small to build a good SMT system. There is some research trying to tackle this problem by using an intermediate *pivot* language, e.g. [1].

It can also happen that the available bitexts do not correspond to the domain for which we want to build a translation system. Many of the available bitexts were produced by multilingual organizations, in particular the European and Canadian Parliament or the United Nations. A particular jargon is often used in these texts, that may not be appropriate for the translation of more general texts. A recent evaluation on automatic translation between European languages has for instance shown that statistical systems perform very well on test data drawn from the European Parliament corpus, i.e. texts of the same type that they were trained on, but their performance can be inferior to rule-based systems for general news data [2].

There are several directions of research to improve the genericity of SMT systems, for instance factored translation model [3], the integration of high quality dictionaries [4] or statistical post-editing of rule-based systems [5, 6]. In this work we investigate whether large-scale unsupervised training is useful to develop a generic SMT system. We define unsupervised training as using the system itself to produce additional bilingual data, i.e. without using a human to perform the translations. The SMT system used to translate the texts was of course itself trained on some bitexts, but these bitexts may be limited in size or little related to the translation task. These resources used to build the initial SMT system are usually not considered as supervision in the framework of unsupervised training applied to *additional data* [7, 8]. An important question is of course to study the success of unsupervised training as a function of the amount of resources used for the initial system. Is it possible to build a powerful SMT using only a very limited amount of initial human-provided resources, i.e. can we replace human-provided bitexts with larger amounts of monolingual data (and its automatic translations) ? Does unsupervised training still work when we already have large amounts of human-translated bitexts ?

In this work we use monolingual data in the target language that may partially cover the same topics than the to text be translated text. It may even contain translations of a fraction of the sentences. It should be noted that these potential partial translations can't be aligned using the standard sentence alignment algorithms. This language model training data can be considered as some form of light supervision and we will therefore use the term *lightly-supervised training* in this work. This can be compared to the research in speech recognition where the same term was used for supervision that either comes from approximate transcriptions of the audio signal (closed captions) or related language model training data. The general idea of our approach is depicted in Figure1.

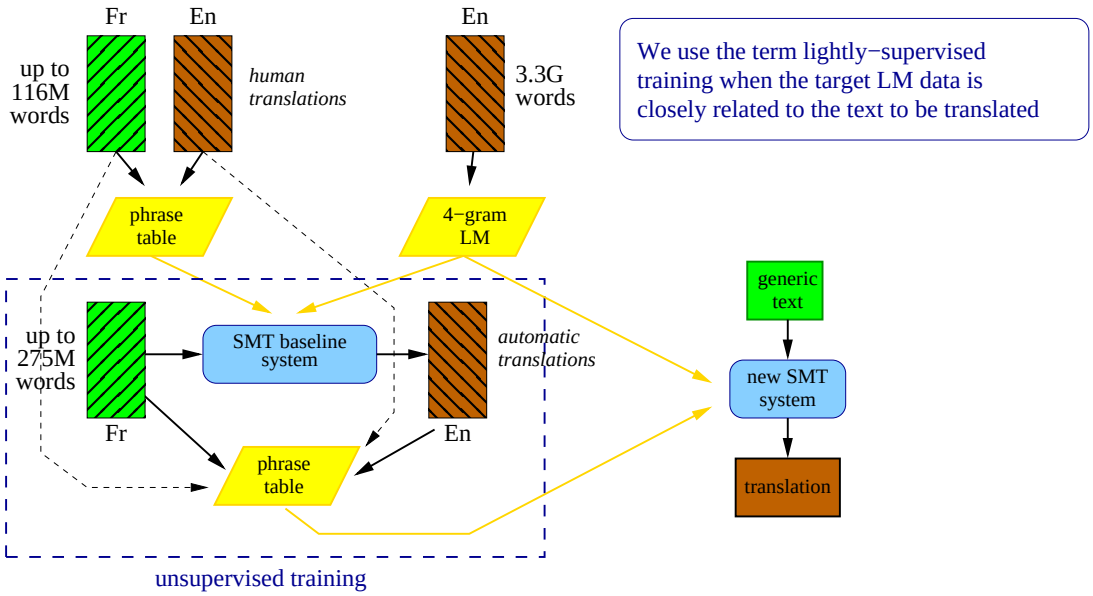

Figure 1: Principle of lightly-supervised training of an SMT system.

Lightly-supervised and unsupervised training has been successfully applied to large vocabulary continuous speech recognition, see for instance [7, 8], but it is not yet widely used in MT. We are only aware of a few pieces of related work. Ueffing et al. used an SMT system 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 used jointly with the generic phrase table [9]. Given the small size of the translated additional data, this technique was presented as domain adaptation rather than unsupervised training of an SMT system. In follow up work, this approach was refined [10], but it was again applied to the test data only. Domain adaptation was also performed simultaneously for the translation, language and reordering model [11].

On the other hand, there is quite some work on the generation of additional bilingual resources to train an SMT system. Munteanu and Marcu proposed an algorithm to automatically detect sentences that are possible translations of each other in large collections of Arabic and English newspaper collections [12]. Their approach does not use an SMT system, but a relatively small word-based bilingual dictionary to translate some of the words of the source sentence. These lexical translations are then used as a query to extract candidate translations using information retrieval techniques. Finally, a maximum entropy classifier is used to select the most promising candidate translations. In another work, a rule-based system was used to generate additional bitexts to train an SMT system [13].

In this paper we investigate whether it is possible to use an SMT system itself to translate several hundred millions of words of monolingual data and to use these automatic translations to improve the SMT system. We concentrate on the translation from French into English. Lightly-supervised training is performed on all the texts from the AFP news agency in LDC's Gigaword collection. This totals several hundreds of millions of words.

In previous work it is mentioned that unsupervised training of SMT systems bears the problem that in principle no new translations can be learned. It is only possible to learn longer phrases of already existing words in the phrase table or to modify the probabilities of existing phrase pairs [9]. The baseline SMT system used in this work includes a large bilingual dictionary. On the one hand, this guarantees that all important words of the source language are in principle known, with the exception of named entities which should be copied over to the target language in most cases anyway. On the other hand, the phrase table entries provided by the dictionary suffer from missing translation probabilities. As an example, the English word *go* may be translated into the French words *aller*, *vais*, *vas*, *allons*, *allez* or *vont*, which should of course not be equally weighted. Lightly-supervised training as used in this paper has the potential to provide better translation probabilities of many of those dictionary words. It should also be beneficial to learn longer phrases that include the dictionary words.

This paper is organized as follows. In the next section we first describe the baseline SMT systems trained on human-provided translations only. The following three sections give details on how large collections of monolingual data were translated, how these texts were filtered and how they were used to train new SMT systems. The paper concludes with a discussion and perspectives of large-scale lightly-supervised training for SMT.

2. Baseline system

The goal of SMT is to produce a target sentence e from a source sentence f. Among all possible target language sentences the one with the highest probability is chosen:

$$e^* = \arg\max_e \Pr(e|f) \quad (1)$$

$$= \arg\max_e \Pr(f|e)\Pr(e) \quad (2)$$

where $\Pr(f|e)$ is the translation model and $\Pr(e)$ is the target language model (LM). This approach is usually referred to as the *noisy source-channel* approach in SMT [14]. Bilingual corpora are needed to train the translation model and monolingual texts to train the target language model.

It is today common practice to use phrases as translation units [15, 16] instead of the original word-based approach. A phrase is defined as a group of source words $\tilde{f}$ that should be translated together into a group of target words $\tilde{e}$. The translation model in phrase-based systems includes the phrase translation probabilities in both directions, i.e. $P(\tilde{e}|\tilde{f})$ and $P(\tilde{f}|\tilde{e})$. The use of a maximum entropy approach simplifies the introduction of several additional models explaining the translation process :

$$e^* = \arg\max Pr(e|f)$$

$$= \arg\max_e \{exp(\sum_i \lambda_i h_i(e,f))\} \quad (3)$$

The feature functions h_i are the system models and the λ_i weights are typically optimized to maximize a scoring function on a development set [17]. In our system fourteen features functions were used, 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.

The system is based on the Moses SMT toolkit [18] and constructed as follows. First, Giza++ is used to perform word alignments in both directions. Second, phrases and lexical reorderings are extracted using the default settings of the Moses SMT toolkit. The 4-gram back-off target LM is trained on the English part of the bitexts and the Gigaword corpus of about 3.2 billion words. The translation model was trained on three parallel corpora:

- the Europarl corpus (40.1M words),
- the news-commentary corpus (1.6M words),
- the Canadian Hansard corpus (72.4M words).

All word counts are given after tokenisation for the French part of the bitexts. In addition, about ten thousand verbs and a hundred thousand nouns from a bilingual dictionary were added to the bitexts. The company SYSTRAN kindly provided this resource. For each verb, we generated all the conjugations in the past, present, future and conditional tense; and for each noun the singular and plural forms were generated. All these forms are provided by the dictionary, including the irregular ones. In total, this resulted in 512k "new sentences" that were directly added to the training data. This has the potential advantage that the dictionary words could improve the alignments of these words when they also appear in the other bitexts. However, one has to be aware that all the translations that appear only in the dictionary will be equally likely which certainly does not correspond to the reality. This is one of our motivations to use lightly-supervised training on large generic corpora. Many of the words in the dictionary are likely to appear in these texts and a better weighting of the corresponding entries in the phrase table can be expected. A more detailed description of this baseline SMT system can be found in [4].

3. Translating large corpora

LDC provides large collections of newspaper texts for language modeling, known as the *Gigaword* corpus.[1] These texts contain about 3.2 billion English and 770 million French words respectively. We identified two news agencies that have provided texts in both languages, namely AFP and APW, and we assume that it is very likely that the texts in both languages cover similar facts. In this study all the texts from AFP were automatically translated with an SMT system from French into English. In our first experiments we only considered the more recent texts (2001–2006, 7.6M sentences and about 275M French words), and included later the texts from the period 1994–1999 (5.3M sentences, 236M words). We also retranslated the Europarl texts from French into English, in order to compare the quality of the automatic translations with the provided human translations.

In the framework of the EuroMatrix project, a test set of general news data was provided for the shared translation task of the third workshop on SMT [2], called *newstest2008* in the following. The size of this corpus amounts to 2051 lines and about 44 thousand words. This data was randomly split into two parts for development and testing. Note that only one reference translation is available. We also noticed several spelling errors in the French source texts, mainly missing accents. These were mostly automatically corrected using the Linux spell checker. This increased the BLEU score by about 1 BLEU point in comparison to the results reported in the official evaluation [2].

The language model interpolation coefficients and the coefficients of the log-linear combination of the feature functions were optimized on this development data before translating the Gigaword corpus. Note that we did use a kind of generic SMT system and that we did not bias the LM towards the text from the Gigaword corpus to be translated,[2] as it was done in some research on lightly-supervised training in automatic speech recognition [8]. We will investigate the benefit of such a bias in future work.

[1] LDC corpora LDC2007T07 (English) and LDC2006T17 (French).

[2] The LM trained on all the texts from AFP has a coefficient of 0.16 in a mixture of 14 language models.

French source text:

- *La paix exige une direction palestinienne nouvelle et différente, afin que puisse naître un Etat palestinien. J'appelle le peuple palestinien à élire de nouveaux dirigeants, des dirigeants qui ne soient pas compromis avec le terrorisme.*
- *M. Arafat, qui s'est juré de faire de l'année 2000 celle de la proclamation d'un Etat palestinien, a mis un point d'honneur à recevoir les six chefs d'Etat présents.*
- *Trois heures après, c'était au tour de la Colombie britannique et de Vancouver de célebrer l'arrivée de l'an nouveau.*
- *"Je m'en étonne et j'évoquerai cette affaire avec le Premier ministre Ehud Barak, car Israël a fait des concessions territoriales aux Palestiniens, et c'est au contraire le moment d'accroître les efforts en faveur de notre sécurité", a déclaré M. Lévy.*

Automatic translations:

- *The peace requires a new and different Palestinian leadership, so that we can create a Palestinian state. I call on the Palestinian people to elect new leaders, leaders not compromised by terrorism.*
- *Mr. Arafat, who has vowed to make the year 2000 the proclamation of a Palestinian state, has made a point of honour to receive the six heads of state present.*
- *Three hours later, it was the turn of the British Columbia and Vancouver célebrer the arrival of the new year.*
- *"I am surprised and I will raise this matter with the Prime Minister Ehud Barak, because Israel has territorial concessions to the Palestinians, and this is the time to increase efforts in favour of our security," said Mr. Lévy.*

Figure 2: Some examples of automatic translations of the Gigaword corpus. Translation errors are underlined. The French word "célebrer" was not translated due to a missing accent. This could be dealt with by performing a spell check prior to translation.

Bitexts	Dict.	Words	Dev	Test
nc	-	1.6M	19.41	19.53
nc	**+**	**2.4M**	**20.44**	**20.18**
nc + ep	-	41.7M	21.96	21.73
nc + ep	+	*43.3M*	*22.27*	*22.35*
nc + hans	-	74.0M	22.06	21.92
nc + hans	+	75.6M	22.04	22.01
nc + ep + hans	-	114M	22.58	22.22
nc + ep + hans	**+**	**116M**	**22.69**	**22.17**

Table 1: BLEU scores of the baseline system using different amounts of human-created bitexts (abbreviations: nc=news-commentary, ep=Europarl, hans=Hansard, dict=dictionary).

The performance of various baseline systems is summarized in Table 1. The word counts are given for the French words after tokenization. The dictionary improves the BLEU score by about 0.6 BLEU points on the test data when used in conjunction with the news-commentary or the Europarl bitexts. The effect is much smaller when the Hansard bitexts are used and the systems using this data are overall slightly worse on the test set, although they do perform better on the development data. We suppose that these bitexts provide too many translations that are specific to bureaucratic texts, but they are not necessarily the best choice for general texts. Two SMT systems were used to translate large amounts of French texts:

1. 2.4M words of bitexts: news-commentary bitexts only and the dictionary (second line in Table 1).

2. 116M words of bitexts: news-commentary, Europarl and Hansard as well as the dictionary. This is the best system that we were able to build using all available human-produced translations (last line in Table 1). We only realized later that better results on the *test set* can be obtained when the Hansard texts are not used (forth line in Table 1).

This corresponds to two extreme cases: a system with very limited resources and a quite large system. In the following we investigate whether lightly-supervised training can be used to improve SMT systems in both conditions.

The phrase table of the big SMT system has about 213M entries and occupies 5.3GB on disk (gzipped compressed). The lexical reordering table is 2GB in size. These models are too big to load them as a whole into memory. Instead, it is common practice to filter them and to only keep the entries that can be applied on the test data. This is however not possible when translating millions of words, even when they are split into smaller parts. Therefore we used the possibility proposed by the Moses decoder to binarize the phrase table and to keep it on the disk. In this representation 46GB of disk space are needed. It is certainly possible to reduce these storage needs by filtering the phrase table in order to suppress unlikely entries, but this was not used in this work. Translation was performed by batches of 200 000 sentences on several machines in parallel. The processing time to translate 275M words amounts to about 1000 hours which corresponds to a translation speed of more than 75 words per second. We anticipate that this could be substantially improved, in particular by using cube-pruning [19] that was recently implemented in the Moses decoder. 100-best lists were generated including the values of the various feature functions and the segmentation information. Figure 2 shows some examples of the automatic translations. We attribute the apparent good quality to the good coverage of our bilingual dictionary and to the quality of the target language model. We plan to make these automatic translations available to the research commu-

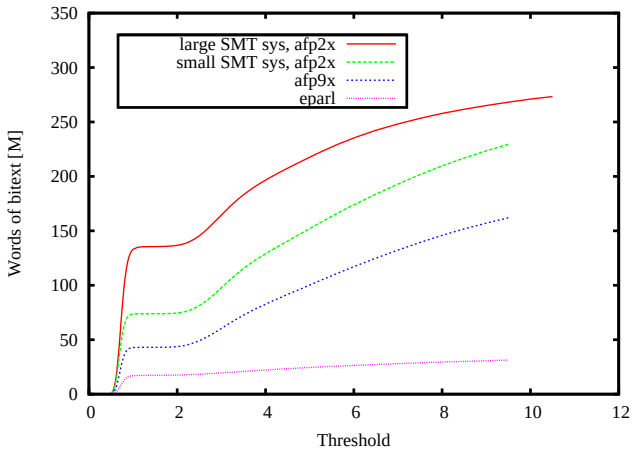

Figure 3: Number of words in the automatically translated texts after filtering with a threshold on the normalized sentence likelihood.

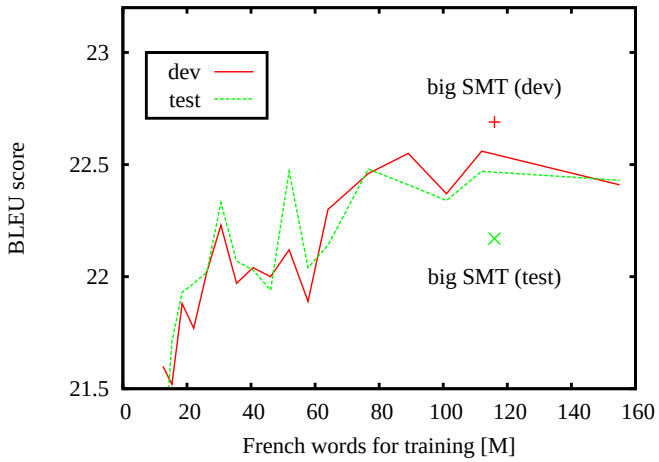

Figure 4: BLEU score when only using the automatic translations of afp2x (translated with the big SMT system).

nity in the future. A binary representation was not necessary for the small SMT system since all the models can be loaded into memory (about 16GB in total).

4. Filtering the automatic translations

Despite the apparent good quality of the automatic translations, it probably makes no sense to use all of them to train an SMT system. Some of the sentences contain rather useless material, like large tables of results of sports events, and others may of course be simply too bad. Therefore it is proposed to filter the translations and to only keep the most reliable ones. This idea was also used when adapting a phrase table to some test data by unsupervised training [9]. In that work, an algorithm was used that explored the n-best list to obtain word-level confidence scores. In this paper we propose a much simpler solution: we directly use the likelihood of the sentence, divided by the number of the words in the hypothesis. Figure 3 shows the numbers of words retained as a function of the threshold on the normalized likelihood of the sentences. All sentences with a value lower than the threshold are kept.

The number of available words increases rapidly until it reaches a plateau for values of the threshold between 1 and 2. This seems to be true for all SMT systems and all corpora. The number of words obtained for a given threshold depends of course on the size of the translated corpora.

5. Using the automatic translations

Several scenarios could be conceived how to use the automatic translations. We could simply add them to the existing bitexts or train instead a separate phrase table. The Moses decoder supports the usage of multiple phrase tables in parallel and independent weights of the feature functions could be learned.

5.1. Using the large SMT baseline system

To start with, we propose to train an SMT system only on the automatic translations of the corpus afp2x (years 2001–2006). This was done with the large SMT baseline system in order to achieve the best possible translation quality. Figure 4 shows the translation performance of such a system as a function of the size of filtered translations used as bitexts. Tuning was performed independently for all experiments reported in this paper.

The BLEU scores on the development set steadily increase with the amount of automatic translations used as bitext (with some noise). Similar observations hold for the performance on the test data. When more than 70M words of automatic translations are used as training bitexts, the

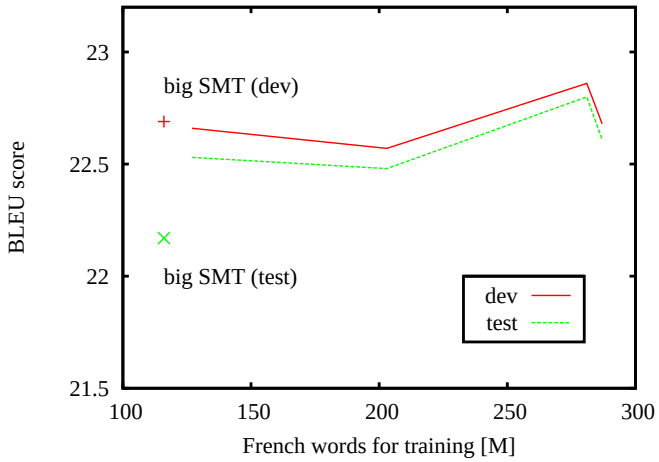

Figure 5: BLEU score when using all human-provided bitexts (114M words) and the automatic translations of afp2x (translated with the big SMT system).

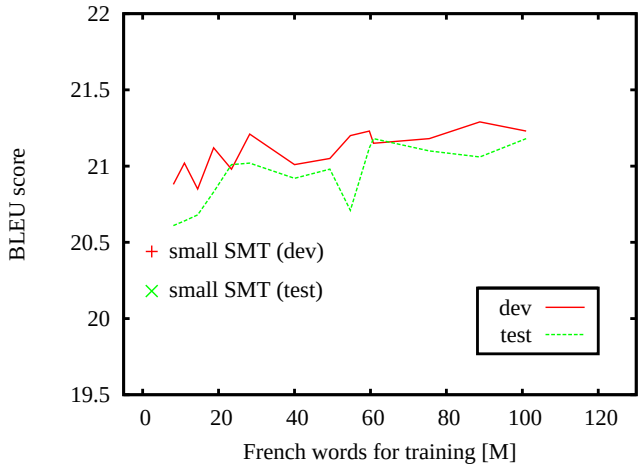

Figure 6: BLEU score when using the news-commentary bitexts and the automatic translations of afp9x (translated with the small SMT system).

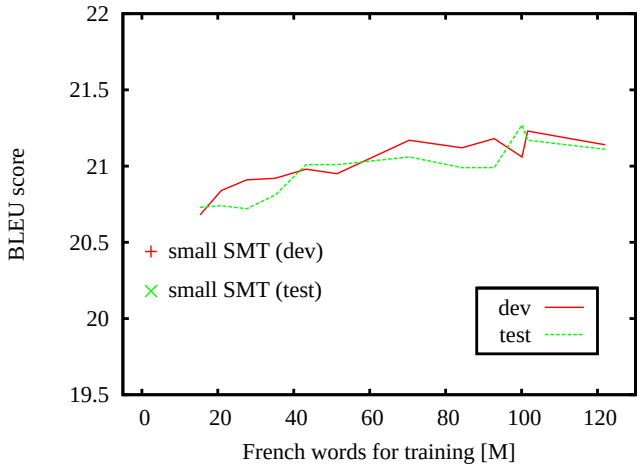

Figure 7: BLEU score when using the news-commentary bitexts and the automatic translations of afp2x (translated with the small SMT system).

performance is even slightly superior to the one of the reference SMT system, which was trained on 114M words of human-translated bitexts and the large dictionary. The system trained on automatic translations only seems to generalize better since the difference between the BLEU score on the development and the test data is smaller than for the reference SMT system.

We also trained systems on all the human-provided translations and the automatically obtained ones. These experiments are shown in figure 5. The BLEU scores on the test data are always superior to the ones obtained with the human-provided translations only, with a fortunate peak when using a total of 280M words of bitexts. This system achieves a BLEU score of 22.80 on the test set, that is 0.6 points higher than the system that was used to translate the monolingual data.

5.2. Using the small SMT baseline system

In a second set of experiments we wanted to investigate whether lightly-supervised training can be used when the initial system is built using a limited amount of human-translated bitexts. In these experiments, only the news-commentary corpus and the bilingual dictionary was used. This system is about two BLEU points worse than the large SMT system (see Table 1). In the following, we will show that half of this loss in performance can be recovered using lightly-supervised training. The Gigaword corpora afp9x and afp2x were translated in two separate experiments, as well as the Europarl corpus (the Europarl bitext are not used in the baseline SMT system).

The BLEU scores on the development and test data are shown in the figures 6 and 7 respectively. Significant improvements with respect to the baseline SMT system were obtained using as little as 10M words of filtered automatic translations. The graphs are quite smooth with an optimum at about 100M filtered words. The best BLEU score on the development data is 21.23. For this setting the BLEU score on the test data increased by 1 BLEU point in comparison to the SMT system that was used to produce the automatic translations. We do not expect that further improvements could be obtained when using more automatically translated texts. Instead we plan to iterate the process, i.e. use this improved system to translate again all the Gigaword corpus, filter the translations, and build a new SMT system. In fact, our experiments in section 5.1 have shown that an SMT system trained on good automatic translations can achieve BLEU scores of more than 22 points on the test set (see figure 4).

Finally, we used the small SMT system to translate the Europarl corpus. This corpus is eventually less relevant to generic news translation due to the particular jargon and vocabulary of the discussions in the European parliament. However, it is interesting to compare the performance of slightly-supervised training with the human-provided translations. In our setting, the correct translations are part of the LM training data, but they have a weight of less than 0.1 in the interpolated LM (since it was optimized on a generic development set). We assume that much better translations could be obtained by using a heavily biased LM.

Figure 8 depicts the BLEU score as a function of the amount of filtered translations that are added to the bitexts. As a comparison, the BLEU scores obtained with the two SMT systems trained on human-provided data are also shown. For training corpora sizes of up to 20M words, the BLEU scores on the development and test sets are similar to those obtained when translating the Gigaword corpora. Beyond this threshold, which roughly corresponds to half of all the data, the automatic translations of the Europarl corpus seem to be too erroneous.

Bitexts Human-provided		Lightly-supervised		Total Words	BLEU score Dev	BLEU score Test	Phrase table Size [#entries]
News+dict	2.4M			2.4M	**20.44**	**20.18**	5M
News+Eparl+dict	43M	-		43.3M	22.17	22.35	83M
News+Eparl+Hans+dict	116M			116M	**22.69**	**22.17**	213M
Translated with the small SMT system:							
News	2.4M	afp9x	28M	2.4M	21.21	21.02	58M
			101M	2.4M	**21.23**	**21.18**	189M
		afp2x	43M	2.4M	20.98	21.01	77M
			102M	2.4M	**21.23**	**21.17**	170M
		Eparl	7M	2.4M	20.78	20.65	17M
			31M	2.4M	21.14	20.86	67M
Translated with the big SMT system:							
-		afp2x	31M	31M	22.23	22.33	55M
			112M	112M	**22.56**	**22.47**	180M
News+Eparl	42M	afp2x	77M	129M	22.65	22.44	203M
	42M		155M	197M	22.53	22.73	320M
News+Eparl+Hans	114M	afp2x	167M	281M	**22.86**	**22.80**	464M

Table 2: Characteristics and result summary of various SMT systems trained on human-provided, on automatic translations or on both.

The different results discussed in this section are summarized in Table 2. Lightly-supervised training achieved improvements in the BLEU score in all settings. The gain on the test set is about 1 BLEU point when only a limited amount of human-provided resources is available (line 5 or 7 in comparison to line 1). To achieve this result about 100M automatically translated and filtered words were added to the bitexts. Slightly lower BLEU scores can be obtained using less bitexts (line 4 or 6 of Table 2). It is important to note that it is not necessary any more to add the dictionary to the bitexts when lightly-supervised training is used.

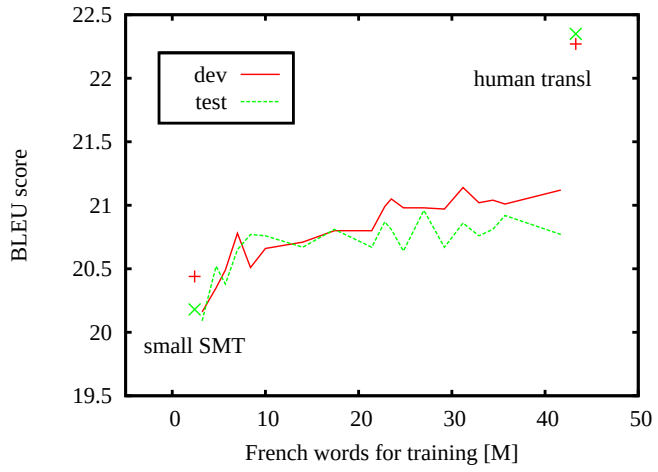

Figure 8: BLEU score when using the news-commentary bitexts and the automatic translations of the Europarl corpus (translated with the small SMT system).

Our best results were obtained by a system trained on 114M words of human-provided translations and 167M words obtained by lightly-supervised training (last line in Table 2). The BLEU score of this pretty big system is 22.80 on the test set, that is 0.6 points higher than the system that was used to translate the monolingual data. For comparable sizes of the parallel training data, the phrase table is always smaller when lightly-supervised training was used. The system trained on 31M words of automatic translations only is particularly interesting since it obtained good BLEU scores with a rather small phrase table (line 10 in Table 2). There also seems to be experimental evidence that lightly-supervised training of SMT systems results in better generalisation behavior since the performance on the development and the test data is in general very similar.

6. Conclusion and discussion

Sentence-aligned bilingual texts are a crucial resource to build SMT systems. For some language pairs bilingual corpora just do not exist, the existing corpora are too small to build a good SMT system or they are not of the same genre or domain. Therefore we studied whether an SMT itself can be used to produce large amounts of automatic translations that can be used as additional training data. We translated up to 275M words of LDC's Gigaword collection from French into English. This is in contrast to previous research that either used rule-based systems to produce additional data [13], or that was based on SMT systems but only applied this idea to very small amounts of data [9, 10, 11]. Those methods of unsupervised training were also called "adaptation" or "self-enhancement" by the respective authors.

We considered two different conditions in this paper. First, we only used a limited amount of human-provided bilingual resources. We started with about 1.6M words of sentence aligned bilingual data and a bilingual dictionary. This system was used to translate large amounts of monolingual data, the automatic translations were filtered and added to the training material. We were able to improve the BLEU score on the test set by about 1 point. The second setup consisted in taking a large, carefully tuned SMT system that was trained on more than 100M words of bilingual sentence-aligned data. Again, the BLEU could be improved by lightly-supervised training.

We believe that these encouraging results open the road to many variations of the proposed method. In this work a generic target LM was used. We will investigate whether a target LM biased towards the topics of the source texts will improve the quality of the automatic translations. Lightly-supervised training could be iterated, alternating the translation of monolingual texts and building an improved systems with these automatic translations. An important step in our procedure is filtering the automatic translations. Currently, we are only using a threshold on the normalized log-likelihood of the sentences. More sophisticated techniques could for instance exploit the score of the individual feature functions, segmentation information, the number of untranslated words or operate on n-best lists. Finally, a limited human analysis has shown that the automatic translations contain several simple errors, like inversion of the adjective-verb order, that could eventually be improved by the integration of various linguistic knowledge sources.

The algorithm proposed in this work is generic and can be applied to any other language pair for which a baseline SMT and large monolingual corpora are available.

7. Acknowledgments

This work has been partially funded by the French Government under the project INSTAR (ANR JCJC06 143038). Some of the baseline SMT systems used in this work were developed in a cooperation between the University of Le Mans and the company SYSTRAN.

8. References

[1] M. Utiyama and H. Isahara, "A comparison of pivot methods for phrase-based statistical machine translation," pp. 484–491, April 2007.

[2] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "Further meta-evaluation of machine translation," pp. 70–106, 2008.

[3] P. Koehn and H. Hoang, "Factored translation models," pp. 868–876, 2007.

[4] H. Schwenk, J.-B. Fouet, and J. Senellart, "First steps towards a general purpose French/English statistical machine translation system," pp. 119–122, 2008.

[5] M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn, "Rule-based translation with statistical phrase-based post-editing," pp. 203–206, 2007.

[6] L. Dugast, J. Senellart, and P. Koehn, "Statistical post-editing on SYSTRAN's rule-based translation system," pp. 220–223, 2007.

[7] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–229, 2002.

[8] L. Nguyen and B. Xiang, "Lightly supervision in acoustic model training," pp. I:185–188, 2004.

[9] N. Ueffing, "Using monolingual source-language data to improve MT performance," pp. 174–181, 2006.

[10] ——, "Transductive learning for statistical machine translation," pp. 25–32, 2007.

[11] B. Chen, M. Zhang, A. Aw, and H. Li, "Exploiting n-best hypotheses for SMT self-enhancement," pp. 157–160, 2008.

[12] D. S. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol. 31, no. 4, pp. 477–504, 2005.

[13] X. Hu, H. Wang, and H. Wu, "Using RBMT systems to produce bilingual corpus for SMT," pp. 287–295, 2007.

[14] P. Brown, S. Della Pietra, V. J. Della Pietra, and R. Mercer, "The mathematics of statistical machine translation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[15] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrased-based machine translation," pp. 127–133, 2003.

[16] F. J. Och and H. Ney, "A systematic comparison of various statistical alignement models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[17] ——, "Discriminative training and maximum entropy models for statistical machine translation," pp. 295–302, 2002.

[18] P. Koehn et al., "Moses: Open source toolkit for statistical machine translation," 2007.

[19] R. C. Moore and C. Quirk, "Faster beam-search decoding for phrasal statistical machine translation," 2007.