LIUM SMT Machine Translation System for WMT 2010

Patrik Lambert, Sadaf Abdul-Rauf and Holger Schwenk
LIUM, University of Le Mans
72085 Le Mans cedex 9, FRANCE
FirstName.LastName@lium.univ-lemans.fr

Abstract

This paper describes the development of French–English and English–French machine translation systems for the 2010 WMT shared task evaluation. These systems were standard phrase-based statistical systems based on the Moses decoder, trained on the provided data only. Most of our efforts were devoted to the choice and extraction of bilingual data used for training. We filtered out some bilingual corpora and pruned the phrase table. We also investigated the impact of adding two types of additional bilingual texts, extracted automatically from the available monolingual data. We first collected bilingual data by performing automatic translations of monolingual texts. The second type of bilingual text was harvested from comparable corpora with Information Retrieval techniques.

1 Introduction

This paper describes the machine translation systems developed by the Computer Science laboratory at the University of Le Mans (LIUM) for the 2010 WMT shared task evaluation. We only considered the translation between French and English (in both directions). The main differences with respect to previous year's system (Schwenk et al., 2009) are as follows: restriction to the data recommended for the workshop, usage of the (filtered) French–English gigaword bitext, pruning of the phrase table, and usage of automatic translations of the monolingual news corpus to improve the translation model. We also used a larger amount of bilingual data extracted from comparable corpora than was done in 2009. These different points are described in the rest of the paper, together with a summary of the experimental results showing the impact of each component.

2 Resources Used

The following sections describe how the resources provided or allowed in the shared task were used to train the translation and language models of the system.

2.1 Bilingual data

Our system was developed in two stages. First, a baseline system was built to generate automatic translations of some of the monolingual data available. These automatic translations may be used directly with the source texts to build additional bitexts, or as queries of an Information Retrieval (IR) system to extract new bitexts from comparable corpora. In a second stage, these additional bilingual data were incorporated to the system (see Section 4 and Tables 1 and 2).

The latest version of the News-Commentary (NC) corpus, of the Europarl (Eparl) corpus (version 5), and of the United Nations (UN) corpus were used. We also took as training data a subset of the French–English Gigaword (10^9) corpus. Since a significant part of the data was crawled from the web, we thought that many sentence pairs may be only approximate translations of each other. We applied a lexical filter to discard them. Furthermore, some sentences of this corpus were extracted from web page menus and are not grammatical. Although we could have used a part of the menu items as a dictionary, for simplicity we applied an n-gram language model (LM) filter to remove all non-grammatical sentences. Thanks to this filter, sentences out of the language model domain (in this case, mainly the news domain), may also have been discarded because they contain many unknown or unfrequent n-grams. The lexical filter was based on the IBM model 1 cost (Brown et al., 1993) of each side of a sentence pair given the other side, normalised with respect to both sentence lengths. This filter

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 127–132,
Uppsala, Sweden, 11-16 July 2010.

was trained on a corpus composed of Eparl, NC, and UN data. The language model filter was an n-gram LM cost of the target sentence (see Section 3), normalised with respect to its length. This filter was trained with all monolingual resources available except the 10^9 data. We generated a first subset, 10^9_1, selecting sentence pairs with a lexical cost inferior to 4 and an LM cost inferior to 2.3. The corpus selected in this way contains 115 million words in the English side (out of 580 million in the original corpus). Close to the evaluation deadline we decided to generate a second corpus (10^9_2) by raising the LM cost threshold to 2.6. The 10^9_2 corpus contains 232 million words on the English side (twice as much as in the 10^9_1 corpus).

In the French side of the bilingual corpora, for the French–English direction only, the contractions 'du' ('of the'), 'au' and 'aux' ('to the' singular and plural) were substituted by their expanded forms ('de le', 'à le' and 'à les').

2.2 Use of Automatic Translations and Comparable corpora

Available human translated bitexts such as the UN corpus seem to be out-of domain for this task. We used two types of automatically extracted resources to adapt our system to the task domain.

First, we generated automatic translations of the French News corpus provided (231M words), and selected the sentences with a normalised translation cost (returned by the decoder) inferior to a threshold. The resulting bitext has no new words in the English side, since all words of the translation output come from the translation model, but it contains new combinations (phrases) of known words, and reinforces the probability of some phrase pairs (Schwenk, 2008).

Second, as in last year's evaluation, we automatically extracted and aligned parallel sentences from comparable in-domain corpora. This year we used the AFP and APW news texts since there are available in the French and English LDC Gigaword corpora. The general architecture of our parallel sentence extraction system is described in detail by Abdul-Rauf and Schwenk (2009). We first translated 91M words from French into English using our first stage SMT system. These English sentences were then used to search for translations in the English AFP and APW texts of the Gigaword corpus using information retrieval techniques. The Lemur toolkit (Ogilvie and Callan, 2001) was used for this purpose. Search was limited to a window of ± 5 days of the date of the French news text. The retrieved candidate sentences were then filtered using the Translation Error Rate (TER) with respect to the automatic translations. In this study, sentences with a TER below 65% for the French–English system and 75% for the English–French system were kept. Sentences with a large length difference (French versus English) or containing a large fraction of numbers were also discarded. By these means, about 15M words of additional bitexts were obtained to include in the French–English system, and 21M words to include in the English–French system. Note that these additional bitexts do not depend on the translation direction. The most suitable amount of additional data was just different in the French–English and English–French translation directions.

2.3 Monolingual data

The French and English target language models were trained on all provided monolingual data. In addition, LDC's Gigaword collection was used for both languages. Data corresponding to the development and test periods were removed from the Gigaword collections.

2.4 Development data

All development was done on *news-test2008*, and *newstest2009* was used as internal test set. For all corpora except the French side of the bitexts used to train the French–English system (see above), the default Moses tokenization was used. However, we added abbreviations for the French tokenizer. All our models are case sensitive and include punctuation. The BLEU scores reported in this paper were calculated with the multi-bleu.perl tool and are case sensitive. The BLEU score was one of metrics with the best correlation with human ratings in last year evaluation (Callison-Burch et al., 2009) for the French–English and English–French directions.

3 Architecture of the SMT system

The goal of statistical machine translation (SMT) is to produce a target sentence e from a source sentence f. It is today common practice to use phrases as translation units (Koehn et al., 2003; Och and Ney, 2003) and a log linear framework in order to introduce several models explaining the

translation process:

$$e^* = \arg\max_e p(e|f) = \arg\max_e \{exp(\sum_i \lambda_i h_i(e,f))\} \quad (1)$$

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 (Och and Ney, 2002). 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 (LM).

The system is based on the Moses SMT toolkit (Koehn et al., 2007) and 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).[1] This speeds up the process and corrects an error of GIZA++ that can appear with rare words.

Phrases and lexical reorderings are extracted using the default settings of the Moses toolkit. The parameters of Moses were tuned on *newstest2008*, using the 'new' MERT tool. We repeated the training process three times, each with a different seed value for the optimisation algorithm. In this way we have an rough idea of the error introduced by the tuning process.

4-gram back-off LMs were used. The word list contains all the words of the bitext used to train the translation model and all words that appear at least ten times in the monolingual corpora. Words of the monolingual corpora containing special characters or sequences of uppercase characters were not included in the word list. Separate LMs were build on each data source with the SRI LM toolkit (Stolcke, 2002) and then linearly interpolated, optimizing the coefficients with an EM procedure. The perplexities of these LMs were 103.4 for French and 149.2 for English.

4 Results and Discussion

The results of our SMT system for the French–English and English–French tasks are summarized in Tables 1 and 2, respectively. The MT metric scores are the average of three optimisations performed with different seeds (see Section 3). The numbers in parentheses are the standard deviation of these three values. The standard deviation gives a lower bound of the significance of the difference between two systems. If the difference between two average scores is less than the sum of the standard deviations, we can say that this difference is not significant. The reverse is not true. Note that most of the improvements shown in the tables are small and not significant. However many of the gains are cumulative and the sum of several small gains makes a significant difference.

Phrase-table Pruning

We tried to prune the phrase-table as proposed by Johnson et. al. (2007), and available in moses ('sigtest-filter'). We used the $\alpha - \epsilon$ filter[2]. As lines 3 and 4 of Table 1, and lines 3 and 4 of Table 2 reveal, in addition to the reduction 43% of the phrase-table, a small gain in BLEU score (0.15 and 0.11 respectively) was obtained with the pruning.

Baseline French–English System

The first section of Table 1 (lines 1 to 5) shows results of the development of the baseline SMT system, used to generate automatic translations. Although being out-of-domain data, the introduction of the UN corpus yields an improvement of one BLEU point with respect to Eparl+NC. Adding the 10^9_1 corpus, we gain 0.7 BLEU point more. Actually, we obtained the same score with the 10^9_1 added directly to Eparl+NC (line 5). However, we choose to include the UN corpus to generate translations to have a larger vocabulary. The system highlighted in bold (line 4) is the one we choose to generate our English translations.

Although no French translations were generated, we did similar experiments in the English–French direction (lines 1 to 4 of Table 2). In this direction, the 10^9_1 corpus is still more valuable than the UN corpus when added to Eparl+NC, but with less difference in terms of BLEU score. In this di-

[1] The source is available at `http://www.cs.cmu.edu/~qing/`

[2] The p-value of two-by-two contingency tables (describing the degree of association between a source and a target phrase) is calculated with Fisher exact test. This probability is interpreted as the probability of observing by chance an association that is at least as strong as the given one, and hence as its significance. An important special case of a table occurs when a phrase pair occurs exactly once in the corpus, and each of the component phrases occurs exactly once in its side of the parallel corpus (1-1-1 phrase pairs). In this case the negative log of the p-value is $\alpha = logN$ (N is number of sentence pairs in the corpus). $\alpha - \epsilon$ is the largest threshold that results in all of the 1-1-1 phrase pairs being included.

rection, we obtain a gain by adding the UN corpus to Eparl+NC+10^9_1.

Filtering the 10^9 Corpus

Lines 5 to 7 of Table 1 show the impact of filtering the 10^9 corpus. The system trained on the full 10^9 corpus added to Eparl+NC achieves a BLEU score of 26.83. Substituting the full 10^9 corpus by 10^9_1 (5 times smaller), i.e. using the first filtering settings, we gain 0.13 BLEU point. Using 10^9_2 instead of 10^9_1, we gain another 0.16 BLEU point, that is 0.3 in total. With respect to not using the 10^9 data at all (as we did last year), we gain 0.8 BLEU point.

Impact of the Additional Bitexts

With the baseline French–English SMT system (see above), we translated the French News corpus to generated an additional bitext (News). We also translated some parts of the French LDC Gigaword corpus, to serve as queries to our IR system (see section 2.2). The resulting additional bitext is referred to as IR. Lines 8 to 13 of Table 1 and lines 6 to 12 of Table 2 summarize the system development including the additional bitexts.

With the News additional bitext added to Eparl+NC, we obtain a system of similar performance as the baseline system used to generate the automatic translations, but with less than 30% of the data. This holds in both translation directions. Adding the News corpus to a larger corpus, such as Eparl+NC+10^9_1, has less impact but still yields some improvement: 0.15 BLEU point in French–English and 0.3 in English–French. Thus, the News bitext translated from French to English may have more impact when translating from English to French than in the opposite direction. Note that the number of additional phrase-table entries per additional running word is twice as high for the News bitext than for the other corpora. For example, with respect to Eparl+NC+UN+10^9_1 (Table 2), Eparl+NC+UN+10^9_1+News has 56M more words and 116M more entries in the phrase-table, thus the ratio is more than 2. For all other corpora, the ratio is equal to 1 or less. This is unexpected, particularly in this case where the News bitext has no new English vocabulary with respect to the Eparl+NC+UN+10^9_1 corpus, from which its English side was generated.

With the IR additional bitext added to Eparl+NC, we obtain a system of similar performance as the system trained on Eparl+NC+UN, while the IR bitext is 10 times smaller than the UN corpus. Added to Eparl+NC+10^9_1+News, the IR bitext allows gains of 0.13 and 0.2 BLEU point respectively in the French–English and English–French directions.

Comparing the systems trained on Eparl+NC+10^9_1 or Eparl+NC+10^9_2 to the systems trained on the same corpora plus News+IR, we can estimate the cumulative impact of the additional bitexts. The gain is around 0.3 BLEU point for French–English and around 0.5 BLEU point for English–French.

Final System

In both translation directions our best system was the one trained on Eparl+NC+10^9_2+News+IR. We further achieved small improvements (0.3 BLEU point) by pruning the phrase-table (as above) and by using a language model with no cut-off together with increasing the beam size and/or the maximum number of translation table entries per input phrase. Note that the English LM with cut-off had a size of 6G, and the one with no cut-off had a size of 29G. It was too much to fit in our 72G machines so we pruned it with the SRILM pruning tool down to a size of 19G. The French LM with cut-off had a size of 2G and the one with no cut-off had a size of 9G. These sizes correspond to the binary format. Taking as example the French–English direction, the running time went from 8600 seconds for the system of line 14 (with a threshold pruning coefficient of 0.4 and a LM with cut-off) to 28200 seconds for the system submitted (with the LM without cut-off pruned by the SRILM tool and a threshold pruning coefficient of 0.00001).

5 Conclusions and Further Work

We presented the development of our machine translation system for the French–English and English–French 2010 WMT shared task. Our system was actually a standard phrase-based SMT system based on the Moses decoder. Its originality mostly lied in the choice and extraction of the training data used.

We decided to use a part of the 10^9 French–English corpus. We found this resource useful, even without filtering. We nevertheless gained 0.3 BLEU point by selecting sentences based on an IBM Model 1 filter and a language model filter.

We pruned the phrase table with the 'sigtest-filter' distributed in Moses, yielding improve-

	Bitext	#Fr Words (M)	P-table size (M)	Mem (G)	news-test2008 BLEU	newstest2009 BLEU
1	Eparl+NC	52	66	19.3	22.80 (0.03)	25.31 (0.2)
2	Eparl+NC+UN	275	250	22.8	23.38 (0.1)	26.30 (0.2)
3	Eparl+NC+UN+10^9_1	406	376	25.1	23.81 (0.05)	27.0 (0.2)
4	**Eparl+NC+UN+10^9_1 pruned**	**406**	**215**	**21.4**	**23.96 (0.1)**	**27.15 (0.18)**
5	Eparl+NC+10^9_1	183	198	22.1	23.83 (0.07)	26.96 (0.04)
6	Eparl+NC+10^9_2	320	319	24.1	23.95 (0.03)	27.12 (0.1)
7	Eparl+NC+10^9	733	580	29.5	23.65 (0.09)	26.83 (0.2)
8	Eparl+NC+News	111	188	19.5	23.46 (0.1)	26.95 (0.2)
9	Eparl+NC+10^9_1+News	242	317	22.5	23.77 (0.04)	27.11 (0.04)
10	Eparl+NC+IR	68	78	19.5	22.97 (0.03)	26.20 (0.1)
11	Eparl+NC+News+IR	127	198	20.1	23.62 (0.01)	27.04 (0.06)
12	Eparl+NC+10^9_1+News+IR	258	327	22.8	23.75 (0.05)	27.24 (0.05)
13	Eparl+NC+10^9_2+News+IR	395	441	24.4	23.87 (0.03)	27.43 (0.08)
14	**Eparl+NC+10^9_2+News+IR pruned** (+larger beam, +no-cutoff LM)	**395**	**285**	**62.5**	**24.04**	**27.72**

Table 1: French–English results: number of French words (in million), number of entries in the phrase-table (in million), memory needed during decoding (in gigabytes) and BLEU scores in the development (news-test2008) and internal test (newstest2009) sets for the different systems developped. The BLEU scores and the number in parentheses are the average and standard deviation over 3 values (see Section 3.)

ments of 0.1 to 0.2 BLEU point for a 43% reduction of the phrase-table size.

We used additional bitexts extracted automatically from the available monolingual corpora. The first type of additional bitext is generated with automatic translations of the monolingual data with a baseline SMT system. The second one is extracted from comparable corpora, with Information Retrieval techniques. With the additional bitexts we gained 0.3 and 0.5 BLEU point for the French–English and English–French systems, respectively.

Next year we want to perform an improved selection of parallel training data with re-sampling techniques. We also want to use a continuous space language model (Schwenk, 2007) in an n-best list rescoring step after decoding. Finally, we plan to train different types of systems (such as a hierarchical SMT system and a Statistical Post-Editing system) and combine their outputs with the MANY open source system combination software (Barrault, 2010).

Acknowledgments

This work has been partially funded by the European Union under the EuroMatrix Plus project – Bringing Machine Translation for European Languages to the User – (http://www.euromatrixplus.net, IST-2007.2.2-FP7-231720).

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece.

Loïc Barrault. 2010. MANY : Open source machine translation system combination. *Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for Machine Translation*, 93:147–155.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the ACL Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece.

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.

	Bitext	#En Words (M)	Phrase-table size (M)	news-test2008 BLEU	newstest2009 BLEU
1	Eparl+NC+UN	242	258	24.21 (0.01)	25.29 (0.12)
2	Eparl+NC+10^9_1	163	203	24.24 (0.06)	25.51 (0.13)
3	Eparl+NC+UN+10^9_1	357	385	24.46 (0.08)	25.73 (0.20)
4	Eparl+NC+UN+10^9_1 pruned	357	221	24.42 (0.1)	25.84 (0.05)
5	Eparl+NC+10^9_2	280	330	24.43 (0.04)	25.68 (0.12)
6	Eparl+NC+News	103	188	24.27 (0.2)	25.70 (0.15)
7	Eparl+NC+10^9_1+News	218	321	24.51 (0.05)	25.83 (0.05)
8	Eparl+NC+UN+10^9_1+News	413	501	24.70 (0.1)	25.86 (0.14)
9	Eparl+NC+IR	69	81	24.14 (0.05)	25.17 (0.2)
10	Eparl+NC+News+IR	124	201	24.32 (0.12)	25.84 (0.17)
11	Eparl+NC+10^9_1+News+IR	239	333	24.54 (0.1)	26.03 (0.15)
12	Eparl+NC+10^9_2+News+IR	356	453	24.68 (0.04)	26.19 (0.05)
13	**Eparl+NC+10^9_2+News+IR pruned** (+larger beam, +no-cutoff LM)	**356**	**293**	**25.06**	**26.53**

Table 2: English–French results: number of English words (in million), number of entries in the phrase-table (in million) and BLEU scores in the development (news-test2008) and internal test (newstest2009) sets for the different systems developped. The BLEU scores and the number in parentheses are the average and standard deviation over 3 values (see Section 3.)

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrased-based machine translation. In *HLT/NACL*, pages 127–133.

Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, demonstration session*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pages 295–302.

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

Paul Ogilvie and Jamie Callan. 2001. Experiments using the Lemur toolkit. In *In Proceedings of the Tenth Text Retrieval Conference (TREC-10)*, pages 103–108.

Holger Schwenk, Sadaf Abdul Rauf, Loïc Barrault, and Jean Senellart. 2009. SMT and SPE machine translation systems for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 130–134, Athens, Greece. Association for Computational Linguistics.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech and Language*, 21:492–518.

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

A. Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, CO.