LIUM's SMT Machine Translation Systems for WMT 2012

Christophe Servan, Patrik Lambert, Anthony Rousseau,
Holger Schwenk and Loïc Barrault
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 statistical machine translation systems for the 2012 WMT shared task evaluation. We developed phrase-based systems based on the Moses decoder, trained on the provided data only. Additionally, new features this year included improved language and translation model adaptation using the cross-entropy score for the corpus selection.

1 Introduction

This paper describes the statistical machine translation systems developed by the Computer Science laboratory at the University of Le Mans (LIUM) for the 2012 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., 2011) are as follows: (i) use of more training data as provided by the organizers and (ii) better selection of the monolingual and parallel data according to the domain, using the cross-entropy difference with respect to in-domain and out-of-domain language models (Moore and Lewis, 2010). We kept some previous features: the improvement of the translation model adaptation by unsupervised training, a parallel corpus retrieved by Information Retrieval (IR) techniques and finally, the rescoring with a continuous space target language model for the translation into French. 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

The latest version of the News-Commentary (NC) corpus and of the Europarl (Eparl) corpus (version 7) were used. We also took as training data a subset of the French–English Gigaword (10^9) corpus. This year we changed the filters applied to select this subset (see Sect. 2.4). We also included in the training data the test sets from previous shared tasks, that we called the ntsXX corpus and which was composed of newstest2008, newstest2009, newssyscomb2009.

2.2 Development data

Development was initially done on *newstest2010*, and *newstest2011* was used as internal test set (Section 3.1). The development and internal test sets were then (Section 4) switched (tuning was done on *newstest2011* and internal evaluation on *newstest2010*). 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 mteval-v13 tool and are case insensitive.

2.3 Use of Automatic Translations

Available human translated bitexts such as the Europarl or 10^9 corpus seem to be out-of domain for this task. We used two types of automatically extracted resources to adapt our system to the domain.

First, we generated automatic translations of the provided monolingual News corpus in French and English, for years 2009, 2010 and 2011, and selected the sentences with a normalised translation cost (returned by the decoder) inferior to a threshold. The resulting bitexts contain no new translations, since all words of the translation output come from the translation model, but they contain new combinations (phrases) of known words, and reinforce the probability of some phrase pairs (Schwenk, 2008). Like last year, we directly used the word-to-word alignments produced by the decoder at the output instead of GIZA's alignments. This speeds-up the procedure and yields the same results in our experiments. A detailed comparison is given in (Lambert et al., 2011).

Second, as in last year's evaluation, we automatically extracted and aligned parallel sentences from comparable in-domain corpora. We used the AFP (Agence France Presse) and APW (Associated Press Worldstream Service) 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 75% were kept. Sentences containing a large fraction of numbers were discarded. By these means, about 27M words of additional bitexts were obtained.

2.4 Domain-based Data selection

Before training the target language models, a text selection has been made using the cross-entropy difference method (Moore and Lewis, 2010). This technique works by computing the difference between two cross-entropy values.

We first score an out-of-domain corpus against a language model trained on a set of in-domain data and compute the cross-entropy for each sentence. Then, we score the same out-of-domain corpus against a language model trained on a random sample of itself, with a size roughly equal to the in-domain corpus. From this point, the difference between in-domain cross-entropy and out-of-domain cross-entropy is computed for each sentence, and these sentences are sorted regarding this score.

By estimating and minimizing on a development set the perplexity of several percentages of the sorted out-of-domain corpus, we can then estimate the theoretical best point of data size for this specific corpus. According the original paper and given our results, this leads to better selection than the simple perplexity sorting (Gao et al., 2002). This way, we can be assured to discard the vast majority of noise in the corpora and to select data well-related to the task.

In this task, the French and English target language models were trained on data selected from all provided monolingual corpora. 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. We had time to apply the domain-based data selection only for French. Thus all data were used for English.

We used this method to filter the French–English 10^9 parallel corpus as well, based on the difference between in-domain cross-entropy and out-of-domain cross-entropy calculated for each sentence of the English side of the corpus. We kept 49 million words (in the English side) to train our models, called 10^9_f.

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. We have build phrase-based systems (Koehn et al., 2003; Och and Ney, 2003), using the standard log linear framework in order to introduce several models explaining the translation process:

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

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, 2003). The phrase-based system uses fourteen features functions, 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 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).[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 using the MERT tool. We repeated the training process three times, each with a different seed value for the optimisation algorithm. In this way we have a 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 on newstest2011 were 119.1 for French and 174.8 for English. In addition, we build a 5-gram continuous space language model for French (Schwenk, 2007). These models were trained on all the available texts using a resampling technique. The continuous space language model is interpolated with the 4-gram back-off model and used to rescore n-best lists. This reduces the perplexity by about 13% relative.

3.1 Number translation

We have also performed some experiments with number translation. English and French do not use the same conventions for integer and decimal numbers. For example, the English decimal number 0.99 is translated in French by 0,99. In the same way, the English integer 32,000 is translated in French by 32 000. It should be possible to perform these modifications by rules.

In this study, we first replaced the numbers by a tag `@@NUM` for integer and `@@DEC` for decimal numbers. Integers in the range 1 to 31 were not replaced since they appear in dates. Then, we created the target language model using the tagged corpora. Table 1 shows results of experiments performed with and without rule-based number translation.

Corpus	NT	BLEU	TER
NC	no	26.57 (0.07)	58.13 (0.06)
NC	yes	**26.84 (0.15)**	**57.71 (0.34)**
Eparl+NC	no	29.28 (0.11)	55.28 (0.13)
Eparl+NC	yes	29.26 (0.10)	55.44 (0.29)

Table 1: Results of the study on number translation (NT) from English to French

We did observe small gains in the translation quality when only the news-commentary bitexts are used, but there were no differences when more training data is available. Due to time constraints, this procedure was not used in the submitted system.

4 Results and Discussion

The results of our SMT systems are summarized in Table 2. The MT metric scores for the development set are the average of three optimisations performed with different seeds (see Section 3). For the test set, they are the average of four values: the three values corresponding to these different optimisations, plus a fourth value obtained by taking as weight for each model, the average of the weights obtained in the three optimisations (**?**). The numbers in parentheses are the standard deviation of these three or four 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.

The results of Table 2 show that adding several adapted corpora (the filtered 10^9 corpus, the syn-

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

Bitext	#Source	newstest2011		newstest2010	
	Words (M)	BLEU	TER	BLEU	TER
Translation : En→Fr					
Eparl+NC	57	30.91 (0.05)	53.61 (0.12)	28.45 (0.08)	56.29 (0.20)
Eparl+NC+ntsXX	58	31.12 (0.08)	53.67 (0.08)	28.49 (0.04)	56.45 (0.12)
Eparl+NC+ntsXX+10^9_f	107	31.67 (0.06)	53.29 (0.03)	29.38 (0.12)	55.45 (0.15)
Eparl+NC+ntsXX+10^9_f+IR	133	32.41 (0.02)	52.20 (0.02)	29.48 (0.11)	55.33 (0.20)
Eparl+NC+ntsXX+10^9_f+news+IR	162	32.26 (0.04)	52.24 (0.12)	29.79 (0.12)	55.04 (0.20)
Translation : Fr→En					
Eparl+NC	64	29.59 (0.12)	51.86 (0.06)	28.12 (0.05)	53.19 (0.06)
Eparl+NC+ntsXX	64	29.59 (0.04)	51.89 (0.14)	28.32 (0.08)	53.22 (0.08)
Eparl+NC+ntsXX+10^9_f	120	30.69 (0.06)	50.77 (0.04)	28.95 (0.14)	52.62 (0.14)
Eparl+NC+ntsXX+10^9_f+IR	149	30.56 (0.02)	50.94 (0.15)	28.67 (0.11)	52.78 (0.06)
Eparl+NC+ntsXX+10^9_f+news+IR	179	30.85 (0.07)	50.72 (0.03)	28.94 (0.05)	52.57 (0.02)

Table 2: English–French and French–English results: number of source words (in million) and scores on the development (newstest2011) and internal test (newstest2010) sets for the different systems developed. The BLEU scores and the number in parentheses are the average and standard deviation over 3 or 4 values when available (see Section 4.)

thetic corpus and the corpus retrieved via IR methods) to the Eparl+NC+ntsXX baseline, a gain of 1.1 BLEU points and 1.4 TER points was achieved for the English–French system.

On the other hand, adding the bitexts extracted from the comparable corpus (IR) does actually hurt the performance of the French–English system: the BLEU score decreases from 28.95 to 28.67 on our internal test set. During the evaluation period, we added all the corpora at once and we observed this only in our analysis after the evaluation.

In both translation directions our best system was the one trained on Eparl+NC+ntsXX+10^9_f+News+IR. Finally, we applied a continuous space language model for the system translating into French.

Acknowledgments

This work has been partially funded by the European Union under the EuroMatrixPlus project ICT-2007.2.2-FP7-231720 and the French government under the ANR project COSMAT ANR-09-CORD-004.

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.

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.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for chinese. In *ACM Transactions on Asian Language Information Processing*.

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

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, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Sixth Workshop on SMT*, pages 284–293.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

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, Patrik Lambert, Loïc Barrault, Christophe Servan, Sadaf Abdul-Rauf, Haithem Afli, and Kashif Shah. 2011. Lium's smt machine translation systems for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 464–469, Edinburgh, Scotland, July. 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.