The RALI Machine Translation System for WMT 2010
Stéphane Huet, Julien Bourdaillet, Alexandre Patry and Philippe Langlais
RALI - Université de Montréal
C.P. 6128, succursale Centre-ville
H3C 3J7, Montréal, Québec, Canada
{huetstep,bourdaij,patryale,felipe}@iro.umontreal.ca

Abstract
We describe our system for the translation task of WMT 2010. This system, developed for the English-French and French-English directions, is based on Moses and was trained using only the resources supplied for the workshop. We report experiments to enhance it with out-of-domain parallel corpora sub-sampling, N-best list post-processing and a French grammatical checker.

1 Introduction
This paper presents the phrase-based machine translation system developed at RALI in order to participate in both the French-English and English-French translation tasks. In these two tasks, we used all the corpora supplied for the constraint data condition apart from the LDC Gigaword corpus.

We describe its different components in Section 2. Section 3 reports our experiments to sub-sample the available out-of-domain corpora in order to adapt the translation models to the news domain. Section 4, dedicated to post-processing, presents how N-best lists are reranked and how the French 1-best output is corrected by a grammatical checker. Section 5 studies how the original source language of news acts upon translation quality. We conclude in Section 6.

2 System Architecture
2.1 Pre-processing
The available corpora were pre-processed using an in-house script that normalizes quotes, dashes, spaces and ligatures. We also reaccentuated French words starting with a capital letter. We significantly cleaned up the parallel Gigaword corpus (noted as gw hereafter), keeping 18.1 M of the original 22.5 M sentence pairs. For example, sentence pairs with numerous numbers, non-alphanumeric characters or words starting with capital letters were removed.

Moreover, training material was tokenized with the tool provided for the workshop and truecased, meaning that the words occurring after a strong punctuation mark were lowercased when they belonged to a dictionary of common all-lowercased forms; the others were left unchanged. In order to reduce the number of words unknown to the translation models, all numbers were serialized, i.e. mapped to a special unique token. The original numbers are then placed back in the translation in the same order as they appear in the source sentence. Since translations are mostly monotonic between French and English, this simple algorithm works well most of the time.

2.2 Language Models
We trained Kneser-Ney discounted 5-gram language models (LMs) on each available corpus using the SRILM toolkit (Stolcke, 2002). These LMs were combined through linear interpolation: first, an out-of-domain LM was built from Europarl, UN and gw; then, this model was combined with the two in-domain LMs trained on news-commentary and news.shuffled, which will be referred to as nc and ns in the remainder of the article. Weights were fixed by optimizing the perplexity of a development corpus made of news-test2008 and news-syscomb2009 texts.

In order to reduce the size of the LMs, we limited the vocabulary of our models to 1 M words for English and French. The words of these vocabularies were selected from the computation of the number of their occurrences using the method proposed by Venkataraman and Wang (2003). The out-of-vocabulary rate measured on news-test2009 and news-test2010 with a so-built vocabulary varies between 0.6 %
and 0.8% for both English and French, while it was between 0.4% and 0.7% before the vocabulary was pruned.

To train the LM on the 48 M-sentence English corpus, 32 Gb RAM were required and up to 16 Gb RAM, for the other corpora. To reduce the memory needs during decoding, LMs were pruned using the SRILM prune option.

2.3 Alignment and Translation Models

All parallel corpora were aligned with Giza++ (Och and Ney, 2003). Our translation models are phrase-based models (PBM) built with Moses (Koehn et al., 2007) with the following non-default settings:

- maximum sentence length of 80 words,
- limit on the number of phrase translations loaded for each phrase fixed to 30.

Weights of LM, phrase table and lexicalized reordering model scores were optimized on the development corpus thanks to the MERT algorithm (Och, 2003).

2.4 Experiments

This section reports experiments done on the news-test2009 corpus for testing various configurations. In these first experiments, we trained LMs and translation models on the Europarl corpus.

Case

We tested two methods to handle case. The first one lowerscases all training data and documents to translate, while the second one normalizes all training data and documents into their natural case. These two methods require a post-processing recapitalization but this last step is more basic for the truecase method. Training models on lowercased material led to a 23.15% case-insensitive BLEU and a 21.61% case-sensitive BLEU; from truecased corpora, we obtained a 23.24% case-insensitive BLEU and a 22.13% case-sensitive BLEU. As truecasing induces an increase of the two metrics, we built all our models in truecase. The results shown in the remainder of this paper are reported in terms of case-insensitive BLEU which showed last year a better correlation with human judgments than case-sensitive BLEU for the two languages we consider (Callison-Burch et al., 2009).

Tokenization

Two tokenizers were tested: one provided for the workshop and another we developed. They differ mainly in the processing of compound words: our in-house tokenizer splits these words (e.g. percentage-wise is turned into percentage-wise), which improves the lexical coverage of the models trained on the corpus. This feature does not exist in the WMT tool. However, using the WMT tokenizer, we measured a 23.24% BLEU, while our in-house tokenizer yielded a lower BLEU of 22.85%. Follow these results prompted us to use the WMT tokenizer.

Serialization

In order to test the effect of serialization, i.e. the mapping of all numbers to a special unique token, we measured the BLEU score obtained by a PBM trained on Europarl for English-French, when numbers are left unchanged (Table 1, line 1) or serialized (line 2). These results exhibit a slight decrease of BLEU when serialization is performed. Moreover, if BLEU is computed using a serialized reference (line 3), which is equivalent to ignoring deserialization errors, a minor gain of BLEU is observed, which validates our recovering method. Since resorting to serialization/deserialization yields comparable performance to a system not using it, while reducing the model’s size, we chose to use it.

|                   | BLEU   |
|-------------------|--------|
| no serialization  | 23.24  |
| corpus serialization | 23.13  |
| corpus and reference serialization | 23.27  |

Table 1: BLEU measured for English-French on news-test2009 when training on Europarl.

LM

Table 2 reports the perplexity measured on news-test2009 for French (column 1) and English (column 3) LMs learned on different corpora and interpolated using the development corpus. We also provide the BLEU score (column 2) for English-French obtained from translation models trained on Europarl and nc. As expected, using in-domain corpora (line 2) for English-French led to better results than using out-of-domain data (line 3). The best perplexities and BLEU score are obtained when LMs trained on all the available corpora are combined (line 4). The last three lines exhibit how LMs perform when they are trained on in-domain corpora without pruning them. While the gzipped 5-gram LM (last line) obtained in
such a manner occupies 1.4 Gb on hard disk, the
gzipped pruned 5-gram LM (line 4) trained using
all corpora occupies 0.9 Gb and yields the same
BLEU score. This last LM was used in all the ex-
periments reported in the subsequent sections.

| corpora                  | Fr ppl | En ppl | BLEU  |
|-------------------------|--------|--------|-------|
| nc                      | 327    | 454    | 22.44 |
| nc + ns                 | 125    | 166    | 25.69 |
| Europarl + UN + Gw      | 156    | 225    | 24.91 |
| all corpora             | 113    | 151    | 26.01 |
| nc + ns (3g, unpruned)  | 138    | -      | 25.32 |
| nc + ns (4g, unpruned)  | 124    | -      | 25.86 |
| nc + ns (5g, unpruned)  | 120    | -      | 26.04 |

Table 2: LMs perplexities and BLEU scores mea-
sured on news-test2009. Translation models
used here were trained on nc and Europarl.

3 Domain adaptation

As the only news parallel corpus provided for
the workshop contains 85k sentence pairs, we
must resort to other parallel out-of-domain cor-
pora in order to build reliable translation models.
If in-domain and out-of-domain LMs are usually
mixed with the well-studied interpolation tech-
niques, training translation models from data of
different domains has received less attention (Fos-
ter and Kuhn, 2007; Bertoldi and Federico, 2009).
Therefore, there is still no widely accepted tech-
nique for this last purpose.

3.1 Effects of the training data size

We investigated how increasing training data acts
upon BLEU score. Table 3 shows a high increase
of 2.7 points w.r.t. the use of nc alone (line 1)
when building the phrase table and the reordering
model from nc and either the 1.7 M-sentence-pair
Europarl (line 2) or a 1.7 M-sentence-pair cor-
pus extracted from the 3 out-of-domain corpora: Eu-
roparl, UN and Gw (line 3). Training a PBM on
merged parallel corpora is not necessarily the best
way to combine data from different domains. We
repeated 20 times nc before adding it to Europarl
so as to have the same amount of out-of-domain
and in-domain material. This method turned out
to be less successful since it led to a minor 0.15
BLEU decrease (line 4) w.r.t. our previous system.

Following the motto “no data is better than more
data”, a PBM was built using all the parallel cor-
pora at our disposal. Since the overall parallel sen-
tences were too numerous for our computational
resources to be simultaneously used, we randomly
split out-of-domain corpora into 3 parts of 8.7 M
sentence pairs each and then combined them with
nc. PBMs were trained on each of these parts
(lines 5 to 7), which yields respectively 0.5 and
0.8 BLEU gain for English-French and French-
English w.r.t. the use of 1.7 M out-of-domain sen-
tence pairs. The more significant improvement no-
ticed for the French-English direction is probably
explained by the fact that the French language is
morphologically richer than English. The 3 PBMs
were then combined by merging the 3 phrase ta-
bles. To do so, the 5 phrase table scores computed
by Moses were mixed using the geometric average
and a 6th score was added, which counts the num-
ber of phrase tables where the given phrase pair
occurs. We ended up with a phrase table contain-
ing 623 M entries, only 9 % and 4 % of them being
in 2 and 3 tables respectively. The resulting phrase
table led to a slight improvement of BLEU scores
(last line) w.r.t. the previous models, except for the
model trained on part 0 for French-English.

3.2 Corpus sub-sampling

Whereas using all corpora improves translation
quality, it requires a huge amount of memory and
disk space. We investigate in this section ways to
select sentence pairs among large out-of-domain
corpora.

Unknown words The main interest of adding new training material relies on the finding of
words missing in the phrase table. According to
this principle, \( nc \) was extended with new sentence pairs containing an unknown word (Table 4, line 2) or a word that belongs to our LM vocabulary and that occurs less than 3 times in the current corpus (line 3). This resulted in adding 400k pairs in the first case and 950k in the second one, with BLEU scores close or even better than those obtained with 1.7 M.

| corpora                        | En→Fr | Fr→En |
|-------------------------------|-------|-------|
| \( nc + 1.7 \text{ M random pairs} \) | 26.02 | 26.68 |
| \( nc + 400k \text{ pairs (occ = 1)} \) | 25.67 | -     |
| \( nc + 950k \text{ pairs (occ = 3)} \) | 26.13 | -     |
| \( nc + \text{Joshua sub-sampling} \) | 26.98 | 27.68 |
| \( nc + \text{IR (1-g q, w/ repet)} \) | 25.81 | -     |
| \( nc + \text{IR (1-g q, no repet)} \) | 26.56 | 27.54 |
| \( nc + \text{IR (1,2-g q, w/ repet)} \) | 26.26 | -     |
| \( nc + \text{IR (1,2-g q, no repet)} \) | 26.53 | -     |
| \( nc + 8.7 \text{ M pairs} \) | 26.68 | 27.65 |
| + IR score (1g q, no repet) | 26.93 | 27.65 |
| 3 large models merged + IR score (1g q, no repet) | 26.86 | 27.56 |
| 26.98 | 27.74 |

Table 4: BLEU measured on \texttt{news-test2009} for English-French and French-English using translation models trained on \( nc \) and a subset of out-of-domain corpora.

**Unknown \( n \)-grams** We applied the sub-sampling method available in the Joshua toolkit (Li et al., 2009). This method adds a new sentence pair when it contains new \( n \)-grams (with \( 1 \leq n \leq 12 \)) occurring less than 20 times in the current corpus, which led us to add 1.5 M pairs for English-French and 1.4 M for French-English. A significant improvement of BLEU is observed using this method (0.8 for English-French and 1.0 for French-English) w.r.t. the use of 1.7 M randomly selected pairs. However, this method has the major drawback of needing to build a new phrase table for each document to translate.

**Information retrieval** Information retrieval (IR) methods have been used in the past to subsample parallel corpora (Hildebrand et al., 2005; Lü et al., 2007). These studies use sentences belonging to the development and test corpora as queries to select the \( k \) most similar source sentences in an indexed parallel corpus. The retrieved sentence pairs constitute a training corpus for the translation models. In order to alleviate the fact that a new PBM has to be learned for each new test corpus, we built queries using sentences contained in the monolingual \( ns \) corpus, leading to the selection of sentence pairs stylistically close to those in the news domain. The source sentences of the three out-of-domain corpora were indexed using Lemur.\(^1\) Two types of queries were built from \( ns \) sentences after removing stop words: the first one is limited to unigrams, the second one contains both unigrams and bigrams, with a weight for bigrams twice as high as for unigrams. The interest of the latter query type is based on the hypothesis that bigrams are more domain-dependent than unigrams. Another choice that needs to be made when using IR methods is concerning the retention of redundant sentences in the final corpus.

Lines 5 to 8 of Table 4 show the results obtained when sentence pairs were gathered up to the size of Europarl, i.e. 1.7 M pairs. 10 sentences were retrieved per query in various configurations: with or without bigrams inside queries, with or without duplicate sentence pairs in the training corpus. Results demonstrate the interest of the approach since the BLEU scores are close to those obtained using the previous tested method based on \( n \)-grams of the test data. Taking bigrams into account does not improve results and adding only once new sentences is more relevant than duplicating them.

Since using all data led to even better performances (see last line of Table 3), we used information provided by the IR method in the PBMs trained on \( nc + 8.7 \text{ M out-of-domain sentence pairs} \) or taking into account all the training material. To this end, we included a new score in the phrase tables which is fixed to 1 for entries that are in the phrase table trained on sentences retrieved with unigram queries without repetition (see line 6 of Table 4), and 0 otherwise. Therefore, this score aims at boosting the weight of phrases that were found in sentences close to the news domain. The results reported in the 4 last lines of Table 4 show minor but consistent gains when adding this score. The outputs of the PBMs trained on all the training corpus and which obtained the best BLEU scores on \texttt{news-test2009} were submitted as contrastive runs. The two first lines of Table 5 report the results on this year’s test data, when the score related to the retrieved corpus is incorporated or not. These results still exhibit a minor improvement when adding this score.

\(^1\)\texttt{www.lemurproject.org}
4 Post-processing

4.1 N-best List Reranking

Our best PBM enhanced by IR methods was employed to generate 500-best lists. These lists were reranked combining the global decoder score with the length ratio between source and target sentences, and the proportions of source sentence n-grams that are in the news monolingual corpora (with $1 \leq n \leq 5$). Weights of these 7 scores are optimized via MERT on news-test2009. Lines 2 and 3 of Table 5 provide the results obtained before and after N-best list reranking. They show a tiny gain for all metrics for English-French, while the results remain constant for French-English. Nevertheless, we decided to use those translations for the French-English task as our primary run.

4.2 Grammatical Checker

PBM outputs contain a significant number of grammatical errors, even when LMs are trained on large data sets. We tested the use of a grammatical checker for the French language: Antidote RX distributed by Druide informatique inc.\textsuperscript{2} This software was applied in a systematic way on the first translation generated after N-best list reranking. Thus, as soon as the software suggests one or several choices that it considers as more correct than the original translation, the first proposal is kept. The checked translation is our first run for English-French.

Antidote RX changed at least one word in 26% of the news-test2010 sentences. The most frequent type of corrections are agreement errors, like in the following example where the agreement between the subject nombre (number) is correctly made with the adjective coupé (cut), thanks to the full syntactic parsing of the French sentence.

\textbf{Source:} [...] the number of revaccinations could then be cut [...] \\
\textbf{Reranking:} [...] le nombre de revaccinations pourrait alors être coupé [...] \\

The example below exhibits a good decision made by the grammatical checker on the mood of the French verb être (to be).

\textbf{Source:} It will be a long time before anything else will be on offer in Iraq. \\
\textbf{Reranking:} Il faudra beaucoup de temps avant que tout le reste sera offert en Irak. \\

\textbf{+Grammatical checker:} Il faudra beaucoup de temps avant que tout le reste soit offert en Irak.

A last interesting type of corrected errors concerns negation. Antidote has indeed the capacity to add the French particle ne when it is missing in the expressions ne ... pas, ne ... plus, aucun ne, personne ne or rien ne. The results obtained using the grammatical checker are reported in the last line of Table 5. The automatic evaluation shows only a minor improvement but we expect the changes induced by this tool to be more significant for human annotators.

5 Effects of the Original Source Language of Articles on Translation

During our experiments, we found that translation quality is highly variable depending on the original source language of the news sentences. This phenomenon is correlated to the previous work of Kurokawa et al. (2009) that showed that whether or not a piece of text is an original or a translation has an impact on translation performance. The main reason that explains our observations is probably that the topics and the vocabulary of news originally expressed in languages other than French and English tend to differ more from those of the training materials used to train PBM models for these two languages. In order to take into account this phenomenon, MERT tuning was repeated for each original source language, using the

\begin{table}[h]
\centering
\begin{tabular}{|l|ccc|ccc|}
\hline & \multicolumn{3}{c|}{En$\rightarrow$Fr} & \multicolumn{3}{c|}{Fr$\rightarrow$En} \\
 & BLEU & BLEU-cased & TER & BLEU & BLEU-cased & TER \\
\hline PBM & 27.5 & 26.5 & 62.2 & 27.8 & 26.9 & 61.2 \\
+IR score & 27.7 & 26.6 & 62.1 & 28.0 & 27.0 & 61.0 \\
+N-best list reranking & 27.9 & 26.8 & 62.1 & 28.0 & 27.0 & 61.2 \\
+grammatical checker & 28.0 & 26.9 & 62.0 & - & - & - \\
\hline
\end{tabular}
\caption{Official results of our system on news-test2010.}
\end{table}
same PBM models trained on all parallel corpora and incorporating an IR score.

Columns 1 and 3 of Table 5 display the BLEU measured using our previous global MERT optimization made on 2553 sentence pairs, while columns 2 and 4 show the results obtained when running MERT on subsets of the development material, made of around 700 sentence pairs each. The BLEU measured on the whole 2010 test set is reported in the last line. As expected, language-dependent MERT tends to increase the LM weight for English and French. However, an absolute 0.35% BLEU decrease is globally observed for English-French using this approach and a 0.21% improvement for French-English.

| MERT   | fr−→En     | fr−→En     |
|--------|-------------|-------------|
| Cz     | 3.95        | 2.14        |
| En     | 30.80       | 29.84       |
| Fr     | 37.59       | 36.96       |
| De     | 16.60       | 16.73       |
| Es     | 24.52       | 24.45       |
| total  | 27.64       | 27.39       |

Table 6: BLEU scores measured on parts of news-test2010 according to the original source language.

6 Conclusion

This paper presented our statistical machine translation system developed for the translation task using Moses. Our submitted runs were generated from models trained on all the corpora made available for the workshop, as this method had provided the best results in our experiments. This system was enhanced using IR methods which exploits news monolingual corpora, N-best list reranking and a French grammatical checker.

This was our first participation where such a huge amount data was involved. Training models on so many sentences is challenging from an engineering point of view and requires important computational resources and storage capacities. The time spent in handling voluminous data prevented us from testing more approaches. We suggest that the next edition of the workshop could integrate a task restraining the number of parameters in the models trained.

References

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In 4th EACL Workshop on Statistical Machine Translation (WMT), Athens, Greece.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In 4th EACL Workshop on Statistical Machine Translation (WMT), Athens, Greece.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In 2nd ACL Workshop on Statistical Machine Translation (WMT), Prague, Czech Republic.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In 10th conference of the European Association for Machine Translation (EAMT), Budapest, Hungary.

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 45th Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume, Prague, Czech Republic.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In 12th Machine Translation Summit, Ottawa, Canada.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, Jonathan Weese, and Omar F. Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In 4th EACL Workshop on Statistical Machine Translation (WMT), Athens, Greece.

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In Join Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In 41st Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan.
Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In 7th International Conference on Spoken Language Processing (ICSLP), Denver, CO, USA.

Armand Venkataraman and Wen Wang. 2003. Techniques for effective vocabulary selection. In 8th European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland.