The LIGA (LIG/LIA) Machine Translation System for WMT 2011

Marion Potet[1], **Raphaël Rubino**[2], **Benjamin Lecouteux**[1], **Stéphane Huet**[2], **Hervé Blanchon**[1], **Laurent Besacier**[1] and **Fabrice Lefèvre**[2]

[1]UJF-Grenoble1, UPMF-Grenoble2
LIG UMR 5217
Grenoble, F-38041, France
FirstName.LastName@imag.fr

[2]Université d'Avignon
LIA-CERI
Avignon, F-84911, France
FirstName.LastName@univ-avignon.fr

Abstract

We describe our system for the news commentary translation task of WMT 2011. The submitted run for the French-English direction is a combination of two MOSES-based systems developed at LIG and LIA laboratories. We report experiments to improve over the standard phrase-based model using statistical post-edition, information retrieval methods to subsample out-of-domain parallel corpora and ROVER to combine n-best list of hypotheses output by different systems.

1 Introduction

This year, LIG and LIA have combined their efforts to produce a joint submission to WMT 2011 for the French-English translation task. Each group started by developing its own solution whilst sharing resources (corpora as provided by the organizers but also aligned data etc) and acquired knowledge (current parameters, effect of the size of n-grams, etc.) with the other. Both LIG and LIA systems are standard phrase-based translation systems based on the MOSES toolkit with appropriate carefully-tuned setups. The final LIGA submission is a combination of the two systems.

We summarize in Section 2 the resources used and the main characteristics of the systems. Sections 3 and 4 describe the specificities and report experiments of resp. the LIG and the LIA system. Section 5 presents the combination of n-best lists hypotheses generated by both systems. Finally, we conclude in Section 6.

2 System overview

2.1 Used data

Globally, our system[1] was built using all the French and English data supplied for the workshop's shared translation task, apart from the Gigaword monolingual corpora released by the LDC. Table 1 sums up the used data and introduces designations that we follow in the remainder of this paper to refer to corpora. Four corpora were used to build translation models: *news-c*, *euro*, *UN* and *giga*, while three others are employed to train monolingual language models (LMs). Three bilingual corpora were devoted to model tuning: *test09* was used for the development of the two seed systems (LIG and LIA), whereas *test08* and *testcomb08* were used to tune the weights for system combination. *test10* was finally put aside to compare internally our methods.

2.2 LIG and LIA system characteristics

Both LIG and LIA systems are phrase-based translation models. All the data were first tokenized with the tokenizer provided for the workshop. Kneser-Ney discounted LMs were built from monolingual corpora using the SRILM toolkit (Stolcke, 2002), while bilingual corpora were aligned at the word-level using GIZA++ (Och and Ney, 2003) or its multi-threaded version MGIZA++ (Gao and Vogel, 2008) for the large corpora *UN* and *giga*. Phrase table and lexicalized reordering models were built with MOSES (Koehn et al., 2007). Finally, 14 features were used in the phrase-based models:

[1]When not specified otherwise "our" system refers to the LIGA system.

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 440–446,
Edinburgh, Scotland, UK, July 30–31, 2011.

CORPORA	DESIGNATION	SIZE (SENTENCES)
English-French Bilingual training		
News Commentary v6	*news-c*	116 k
Europarl v6	*euro*	1.8 M
United Nation corpus	*UN*	12 M
10^9 corpus	*giga*	23 M
English Monolingual training		
News Commentary v6	*mono-news-c*	181 k
Shuffled News Crawl corpus (from 2007 to 2011)	*news-s*	25 M
Europarl v6	*mono-euro*	1.8 M
Development		
newstest2008	*test08*	2,051
newssyscomb2009	*testcomb09*	502
newstest2009	*test09*	2,525
Test		
newstest2010	*test10*	2,489

Table 1: Used corpora

- 5 translation model scores,
- 1 distance-based reordering score,
- 6 lexicalized reordering score,
- 1 LM score and
- 1 word penalty score.

The score weights were optimized on the *test09* corpus according to the BLEU score with the MERT method (Och, 2003). The experiments led specifically with either LIG or LIA system are respectively described in Sections 3 and 4. Unless otherwise indicated, all the evaluations were performed using case-insensitive BLEU and were computed with the `mteval-v13a.pl` script provided by NIST. Table 2 summarizes the differences between the final configuration of the systems.

3 The LIG machine translation system

LIG participated for the second time to the WMT shared news translation task for the French-English language pair.

3.1 Pre-processing

Training data were first lowercased with the PERL script provided for the campaign. They were also processed in order to normalize a special French form (named euphonious "t") as described in (Potet et al., 2010).

The baseline system was built using a 4-gram LM trained on the monolingual corpora provided last year and translation models trained on *news-c* and *euro* (Table 3, System 1). A significant improvement in terms of BLEU is obtained when taking into account a third corpus, *UN*, to build translation models (System 2). The next section describes the LMs that were trained using the monolingual data provided this year.

3.2 Language model training

Target LMs are standard 4-gram models trained on the provided monolingual corpus (*mono-news-c*, *mono-euro* and *news-s*). We decided to test two different n-gram cut-off settings. The fist set has low cut-offs: 1-2-3-3 (respectively for 1-gram, 2-gram, 3-gram and 4-gram counts), whereas the second one (LM_2) is more aggressive: 1-5-7-7. Experiment results (Table 3, Systems 3 and 4) show that resorting to LM_2 leads to an improvement of BLEU with respect to LM_1. LM_2 was therefore used in the subsequent experiments.

FEATURES	LIG SYSTEM	LIA SYSTEM
Pre-processing	Text lowercased Normalization of French euphonious 't'	Text truecased Reaccentuation of French words starting with a capital letter
LM	Training on *mono-news-c*, *news-s* and *mono-euro* 4-gram models	Training on *mono-news-c* and *news-s* 5-gram models
Translation model	Training on *news-c*, *euro* and *UN* Phrase table filtering Use of *-monotone-at-punctuation* option	Training on 10 M sentence pairs selected in *news-c*, *euro*, *UN* and *giga*

Table 2: Distinct features between final configurations retained for the LIG and LIA systems

3.3 Translation model training

Translation models were trained from the parallel corpora *news-c*, *euro* and *UN*. Data were aligned at the word-level and then used to build standard phrase-based translation models. We filtered the obtained phrase table using the method described in (Johnson et al., 2007). Since this technique drastically reduces the size of the phrase table, while not degrading (and even slightly improving) the results on the development and test corpora (System 6), we decided to employ filtered phrase tables in the final configuration of the LIG system.

3.4 Tuning

For decoding, the system uses a log-linear combination of translation model scores with the LM log-probability. We prevent phrase reordering over punctuation using the MOSES option *-monotone-at-punctuation*. As the system can be beforehand tuned by adjusting the log-linear combination weights on a development corpus, we used the MERT method (System 5). Optimizing weights according to BLEU leads to an improvement with respect to the system with MOSES default value weights (System 5 *vs* System 4).

3.5 Post-processing

We also investigated the interest of a statistical post-editor (SPE) to improve translation hypotheses. About 9,000 sentences extracted from the news domain test corpora of the 2007–2009 WMT translation tasks were automatically translated by a system very similar to that described in (Potet et al., 2010), then manually post-edited. Manual corrections of translations were performed by means of the crowd-sourcing platform AMAZON MECHANICAL TURK[2] ($0.15/sent.). These collected data make a parallel corpus whose source part is MT output and target part is the human post-edited version of MT output. This are used to train a phrase-based SMT (with Moses without the tuning step) that automatically post-edit the MT output. That aims at learning how to correct translation hypotheses. System 7 obtained when post-processing MT 1-best output shows a slight improvement. However, SPE was not used in the final LIG system since we lacked time to apply SPE on the N-best hypotheses for the development and test corpora (the N-best being necessary for combination of LIG and LIA systems). Ths LIGA submission is thus a constrained one.

3.6 Recasing

We trained a phrase-based recaser model on the *news-s* corpus using the provided MOSES scripts and applied it to uppercase translation outputs. A common and expected loss of around 1.5 case-sensitive BLEU points was observed on the test corpus (*news10*) after applying this recaser (System 7) with respect to the score case-insensitive BLEU previously measured.

[2] `http://www.mturk.com/mturk/welcome`

♯	SYSTEM DESCRIPTION	BLEU SCORE *test09*	*test10*
1	Training: *euro+news-c*	24.89	26.01
2	**Training:** *euro+news-c+UN*	25.44	26.43
3	2 + LM_1	24.81	27.19
4	2 + $\mathbf{LM_2}$	25.37	27.25
5	4 + **MERT** on *test09*	26.83	27.53
6	5 + **phrase-table filtering**	27.09	**27.64**
7	6 + SPE	27.53	27.74
8	6 + recaser	24.95	26.07

Table 3: Incremental improvement of the LIG system in terms of case-insensitive BLEU (%), except for line 8 where case-sensitive BLEU (%) are reported

4 The LIA machine translation system

This section describes the particularities of the MT system which was built at the LIA for its first participation to WMT.

4.1 System description

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 crawled parallel *giga* corpus, keeping 19.3 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. The whole training material is truecased, meaning that the words occuring after a strong punctuation mark were lowercased when they belonged to a dictionary of common all-lowercased forms; the others were left unchanged.

The training of a 5-gram English LM was restrained to the news corpora *mono-news-c* and *news-s* that we consider large enough to ignore other data. In order to reduce the size of the LM, we first limited the vocabulary of our model to a 1 M word vocabulary taking the most frequent words in the news corpora. We also resorted to cut-offs to discard infrequent n-grams (2-2-3-5 thresholds on 2- to 5-gram counts) and uses the SRILM option `prune`, which allowed us to train the LM on large data with 32 Gb RAM.

Our translation models are phrase-based models (PBMs) built with MOSES 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.

Besides the size of used data, we experimented with two advanced features made available for MOSES. Firstly, we filtered phrase tables using the default setting `-l a+e -n 30`. This dramatically reduced phrase tables by dividing their size by a factor of 5 but did not improve our best configuration from the BLEU score perspective (Table 4, line 1); the method was therefore not kept in the LIA system. Secondly, we introduced reordering constraints in order to consider quoted material as a block. This method is particularly useful when citations included in sentences have to be translated. Two configurations were tested: *zone* markups inclusion around quotes and *wall* markups inclusion within *zone* markups. However, the measured gains were finally too marginal to include the method in the final system.

4.2 Parallel corpus subsampling

As the only news parallel corpus provided for the workshop contains 116 k sentence pairs, we must resort to parallel out-of-domain corpora in order to build reliable translation models. Information retrieval (IR) methods have been used in the past to subsample parallel corpora. For example, Hildebrand et al. (2005) used 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 constituted a training corpus for the translation models.

The RALI submission for WMT10 proposed a similar approach that builds queries from the monolingual news corpus in order to select sentence pairs stylistically close to the news domain (Huet et al., 2010). This method has the major interest that it does not require to build a new training parallel corpus for each news data set to translate. Following the best configuration tested in (Huet et al.,

2010), we index the three out-of-domain corpora using LEMUR[3], and build queries from English *news-s* sentences where stop words are removed. The 10 top sentence pairs retrieved per query are selected and added to the new training corpus if they are not redundant with a sentence pair already collected. The process is repeated until the training parallel corpus reaches a threshold over the number of retrieved pairs.

Table 4 reports BLEU scores obtained with the LIA system using the in-domain corpus *news-c* and various amounts of out-of-domain data. MERT was re-run for each set of training data. The first four lines display results obtained with the same number of sentence pairs, which corresponds to the size of *news-c* appended to *euro*. The experiments show that using *euro* instead of the first sentences of *UN* and *giga* significantly improves BLEU scores, which indicates the better adequacy of *euro* with respect to the *test10* corpus. The use of the IR method to select sentences from *euro*, *UN* and *giga* leads to a similar BLEU score to the one obtained with *euro*. The increase of the collected pairs up to 3 M pairs generates a significant improvement of 0.9 BLEU point. A further rise of the amount of collected pairs does not introduce a major gain since retrieving 10 M sentence pairs only augments BLEU from 29.1 to 29.3. This last configuration which leads to the best BLEU was used to build the final LIA system. Let us note that 2 M, 3 M and 15 M queries were required to respectively obtain 3 M, 5 M and 10 M sentence pairs because of the removal of redundant sentences in the increased corpus.

For a matter of comparison, a system was also built taking into account all the training material, i.e. 37 M sentence pairs[4]. This last system is outperformed by our best system built with IR and has finally close performance to the one obtained with *news-c+euro* relatively to the quantity of used data.

USED PARALLEL CORPORA	FILTERING	
	without	with
news-c + *euro* (1.77 M)	28.1	28.0
news-c + 1.77 M of *UN*	27.2	-
news-c + 1.77 M of *giga*	27.1	-
news-c + 1.77 M with IR	28.2	-
news-c + 3 M with IR	29.1	29.0
news-c + 5 M with IR	28.8	-
news-c **+ 10 M with IR**	**29.3**	29.2
All data	28.9	29.0

Table 4: BLEU (%) on test10 measured with the LIA system using different training parallel corpora

5 The system combination

System combination is based on the 500-best outputs generated by the LIA and the LIG systems. They both used the MOSES option `distinct`, ensuring that the hypotheses produced for a given sentence are different inside an N-best list. Each N-best list is associated with a set of 14 scores and combined in several steps.

The first step takes as input lowercased 500-best lists, since preliminary experiments have shown a better behavior using only lowercased output (with cased output, combination presents some degradations). The score combination weights are optimized on the development corpus, in order to maximize the BLEU score at the sentence level when N-best lists are reordered according to the 14 available scores. To this end, we resorted to the SRILM `nbest-optimize` tool to do a simplex-based Amoeba search (Press et al., 1988) on the error function with multiple restarts to avoid local minima.

Once the optimized feature weights are computed independently for each system, N-best lists are turned into confusion networks (Mangu et al., 2000). The 14 features are used to compute posteriors relatively to all the hypotheses in the N-best list. Confusion networks are computed for each sentence and for each system. In Table 5 we present the ROVER (Fiscus, 1997) results for the LIA and LIG confusion networks (LIA CNC and LIG CNC). Then, both confusion networks computed for each sentence are merged into a single one. A ROVER is applied on the combined confusion network and generates a lowercased 1-best.

The final step aims at producing cased hypotheses. The LIA system built from truecased corpora achieved significantly higher performance than the

[3] `www.lemurproject.org`

[4] For this experiment, the data were split into three parts to build independent alignment models: *news-c+euro*, *UN* and *giga*, and they were joined afterwards to build translation models.

		LIG	LIA	LIG CNC	LIA CNC	LIG+LIA
case-insensitive BLEU	*test10*	27.6	29.3	28.1	29.4	29.7
	test11	28.5	29.4	28.5	29.3	29.9
case-sensitive BLEU	*test10*	26.1	28.4	27.0	28.4	28.7
	test11	26.9	28.4	27.5	28.4	28.8

Table 5: Performance measured before and after combining systems

LIG system trained on lowercased corpora (Table 5, two last lines). In order to get an improvement when combining the outputs, we had to adopt the following strategy. The 500-best truecased outputs of the LIA system are first merged in a word graph (and not a mesh lattice). Then, the lowercased 1-best previously obtained with ROVER is aligned with the graph in order to find the closest existing path, which is equivalent to matching an oracle with the graph. This method allows for several benefits. The new hypothesis is based on a "true" decoding pass generated by a truecased system and discarded marginal hypotheses. Moreover, the selected path offers a better BLEU score than the initial hypothesis with and without case. This method is better than the one which consists of applying the LIG recaser (section 3.6) on the combined (un-cased) hypothesis.

The new recased one-best hypothesis is then used as the final submission for WMT. Our combination approach improves on *test11* the best single system by 0.5 case-insensitive BLEU point and by 0.4 case-sensitive BLEU (Table 5). However, it also introduces some mistakes by duplicating in particular some segments. We plan to apply rules at the segment level in order to reduce these artifacts.

6 Conclusion

This paper presented two statistical machine translation systems developed at different sites using MOSES and the combination of these systems. The LIGA submission presented this year was ranked among the best MT system for the French-English direction. This campaign was the first shot for LIA and the second for LIG. Beside following the traditional pipeline for building a phrase-based translation system, each individual system led to specific works: LIG worked on using SPE as post-treatment, LIA focused on extracting useful data from large-sized corpora. And their combination implied to address the interesting issue of matching results from systems with different casing approaches.

WMT is a great opportunity to chase after performance and joining our efforts has allowed to save considerable amount of time for data preparation and tuning choices (even when final decisions were different among systems), yet obtaining very competitive results. This year, our goal was to develop state-of-the-art systems so as to investigate new approaches for related topics such as translation with human-in-the-loop or multilingual interaction systems (e.g. vocal telephone information-query dialogue systems in multiple languages or language portability of such systems).

References

Jonathan G. Fiscus. 1997. A post-processing system to yield reduced word error rates:recognizer output voting error reduction (ROVER). In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–354, Santa Barbara, CA, USA.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Proceedings of the ACL Workshop: Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, OH, USA.

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 *Proceedings of the 10th conference of the European Association for Machine Translation (EAMT)*, Budapest, Hungary.

Stéphane Huet, Julien Bourdaillet, Alexandre Patry, and Philippe Langlais. 2010. The RALI machine translation system for WMT 2010. In *Proceedings of the ACL Joint 5th Workshop on Statistical Machine Translation and Metrics (WMT)*, Uppsala, Sweden.

Howard Johnson, Joel Martin, George Foster, and Roland

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

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 *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume*, pages 177–180, Prague, Czech Republic, June.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400.

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 *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, Sapporo, Japan.

Marion Potet, Laurent Besacier, and Hervé Blanchon. 2010. The LIG machine translation for WMT 2010. In *Proceedings of the ACL Joint 5th Workshop on Statistical Machine Translation and Metrics (WMT)*, Uppsala, Sweden.

William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1988. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Andreas Stolcke. 2002. SRILM — an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, USA.