LIMSI's statistical translation systems for WMT'08

Daniel Déchelotte, Gilles Adda, Alexandre Allauzen, Hélène Bonneau-Maynard, Olivier Galibert, Jean-Luc Gauvain, Philippe Langlais* and **François Yvon**
LIMSI/CNRS
firstname.lastname@limsi.fr

Abstract

This paper describes our statistical machine translation systems based on the Moses toolkit for the WMT08 shared task. We address the Europarl and News conditions for the following language pairs: English with French, German and Spanish. For Europarl, n-best rescoring is performed using an enhanced n-gram or a neuronal language model; for the News condition, language models incorporate extra training data. We also report unconvincing results of experiments with factored models.

1 Introduction

This paper describes our statistical machine translation systems based on the Moses toolkit for the WMT 08 shared task. We address the Europarl and News conditions for the following language pairs: English with French, German and Spanish. For Europarl, n-best rescoring is performed using an enhanced n-gram or a neuronal language model, and for the News condition, language models are trained with extra training data. We also report unconvincing results of experiments with factored models.

2 Base System architecture

LIMSI took part in the evaluations on Europarl data and on News data, translating French, German and Spanish from and to English, amounting a total of twelve evaluation conditions. Figure 1 presents the generic overall architecture of LIMSI's translation systems. They are fairly standard phrase-based translation systems (Och and Ney, 2004; Koehn et al., 2003) and use Moses (Koehn et al., 2007) to search for the best target sentence. The search uses the following models: a phrase table, providing 4 scores and a phrase penalty, a *lexicalized* reordering model (7 scores), a language model score and a word penalty. These fourteen scores are weighted and linearly combined (Och and Ney, 2002; Och, 2003); their respective weights are learned on development data so as to maximize the BLEU score. In the following, we detail several aspects of our systems.

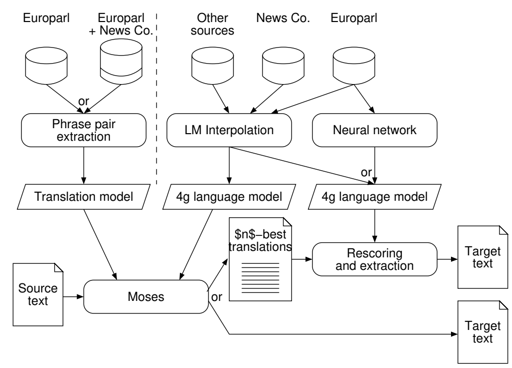

Figure 1: Generic architecture of LIMSI's SMT systems. Depending on the condition, the decoder generates either the final output or n-best lists. In the latter case, the rescoring incorporates the same translation features, except for a better target language model (see text).

2.1 Translation models

The translation models deployed in our systems for the europarl condition were trained on the provided Europarl parallel data only. For the news condition, they were trained on the Europarl data merged with

*Univ. Montréal, felipe@iro.umontreal.ca

Proceedings of the Third Workshop on Statistical Machine Translation, pages 107–110,
Columbus, June 2008.

the news-commentary parallel data, as depicted on Figure 1. This setup was found to be more favorable than training on Europarl data only (for obvious mismatching domain reasons) and than training on news-commentary data only, most probably because of a lack of coverage. Another, alternative way of benefitting from the coverage of the Europarl corpus *and* the relevance of the news-commentary corpus is to use two phrase-tables *in parallel*, an interesting feature of Moses. (Koehn and Schroeder, 2007) found that this was the best way to "adapt" a translation system to the news-commentary task. These results are corroborated in (Déchelotte, 2007)[1] , which adapts a "European Parliament" system using a "European and Spanish Parliaments" development set. However, we were not able to reproduce those findings for this evaluation. This might be caused by the increase of the number of feature functions, from 14 to 26, due to the duplication of the phrase table and the lexicalized reordering model.

2.2 Language Models

2.2.1 Europarl language models

The training of Europarl language models (LMs) was rather conventional: for all languages used in our systems, we used a 4-gram LM based on the entire Europarl vocabulary and trained only on the available Europarl training data. For French, for instance, this yielded a model with a 0.2 out-of-vocabulary (OOV) rate on our LM development set, and a perplexity of 44.9 on the development data. For French also, a more accurate n-gram LM was used to rescore the first pass translation; this larger model includes both Europarl and giga word corpus of newswire text, lowering the perplexity to 41.9 on the development data.

2.2.2 News language models

For this condition, we took advantage of the a priori information that the test text would be of newspaper/newswire genre and from the November-december 2007 period. We consequently built much larger LMs for translating both to French and to English, and optimized their combination on appropriate source of data. For French, we interpolated five different LMs trained on corpus containing respectively newspapers, newswire, news commentary and Europarl data, and tuned their combination with text downloaded from the Internet. Our best LM had an OOV rate of about 2.1% and a perplexity of 111.26 on the testset. English LMs were built in a similar manner, our largest model combining 4 LMs from various sources, which, altogether, represent about 850M words. Its perplexity on the 2008 test set was approximately 160, with an OOV rate of 2.7%.

[1](Déchelotte, 2007) further found that giving an increased weight to the small in-domain data could out-perform the setup with two phrase-tables in parallel. We haven't evaluated this idea for this evaluation.

2.2.3 Neural network language models

Neural-Network (NN) based continuous space LMs similar to the ones in (Schwenk, 2007) were also trained on Europarl data. These networks compute the probabilities of all the words in a 8192 word output vocabulary given a context in a larger, 65000-word vocabulary. Each word in the context is first associated with a numerical vector of dimension 500 by the input layer. The activity of the 500 neurons in the hidden layer is computed as the hyperbolic tangent of the weighted sum of these vectors, projecting the context into a $[-1, 1]$ hypercube of dimension 500. Final projection on a set of 8192 output neurons yields the final probabilities through a softmax-ed, weighted sum of the coordinates in the hypercube. The final NN-based model is interpolated with the main LM model in a 0.4-0.6 ratio, and yields a perplexity reduction of 9% relative with respect to the n-gram LM on development data.

2.3 Tuning procedure

We use MERT, distributed with the Moses decoder, to tune the first pass of the system. The weights were adjusted to maximize BLEU on the development data. For the baseline system, a dozen Moses runs are necessary for each MERT optimization, and several optimization runs were started and compared during the system's development. Tuning was performed using dev2006 for the Europarl task and on News commentary dev2007 for the news task.

2.4 Rescoring and post processing

For the Europarl condition, distinct 100 best translations from Moses were rescored with improved LMs: when translating to French, we used the French model described in section 2.2.1; when

	Es-En	En-Es	Fr-En	En-Fr
baseline	32.21	31.62	32.41	29.31
Limsi	32.49	31.23	32.62	30.27

Table 1: Comparison of two tokenization policies
All results on Europarl test2007

	CI system	CS system
En→Fr	27.23	27.55
Fr→En	30.96	30.98

Table 2: Effect of training on true case texts, for English to French (case INsensitive BLEU scores, untuned systems, results on test2006 dataset)

translating to English, we used the neuronal LM described in section 2.2.3.

For all the "lowcase" systems (see below), recaseing was finally performed using our own recaseing tool. Case is restored by creating a word graph allowing all possible forms of caseing for each word and each component of a compound word. This word graph is then decoded using a cased 4-gram LM to obtain the most likely form. In a final step, OOV words (with respect to the source language word list) are recased to match their original form.

3 Experiments with the base system

3.1 Word tokenization and case

We developed our own tokenizer for English, French and Spanish, and used the baseline tokenizer for German. Experiments on the 2007 test dataset for Europarl task show the impact of the tokenization on the BLEU scores, with 3-gram LMs. Results are always improved with our own tokenizer, except for English to Spanish (Table 1).

Our systems were initially trained on lowercase texts, similarly to the proposed baseline system. However, training on true case texts proved beneficial when translating from English to French, even when scoring in a case insensitive manner. Table 2 shows an approximate gain of 0.3 BLEU for that direction, and no impact on French to English performance. Our English-French systems are therefore case sensitive.

3.2 Language Models

For Europarl, we experimented with LMs of increasing orders: we found that using a 5-gram LM only yields an insignificant improvement over a 4-gram LM. As a result, we used 4-gram LMs for all our first pass decodings. For the second pass, the use of the Neural Network LMs, if used with an appropriate (tuned) weight, yields a small, yet consistent improvement of BLEU for all pairs.

Performance on the news task are harder to analyze, due to the lack of development data. Throwing in large set of in-domain data was obviously helpful, even though we are currently unable to adequately measure this effect.

4 Experiments with factored models

Even though these models were not used in our submissions, we feel it useful to comment here our (negative) experiments with factored models.

4.1 Overview

In this work, factored models (Koehn and Hoang, 2007) are experimented with three factors : the surface form, the lemma and the part of speech (POS). The translation process is composed of different mapping steps, which either translate input factors into output factors, or generate additional output factors from existing output factors. In this work, four mapping steps are used with two decoding paths. The first path corresponds to the standard and direct mapping of surface forms. The second decoding path consists in two translation steps for respectively POS tag and the lemmas, followed by a generation step which produces the surface form given the POS-lemma couple. The system also includes three reordering models.

4.2 Training

Factored models have been built to translate from English to French for the *news* task. To estimate the phrase and generation tables, the training texts are first processed in order to compute the lemmas and POS information. The English texts are tagged and lemmatized using the English version of the Treetagger[2]. For French, POS-tagging is carried out with a French version of the Brill's tagger trained

[2]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

on the MULTITAG corpus (Allauzen and Bonneau-Maynard, 2008). Lemmatization is performed with a French version of the Treetagger.

Three phrase tables are estimated with the Moses utilities, one per factor. For the surface forms, the parallel corpus is the concatenation of the official training data for the tasks Europarl and News commentary, whereas only the parallel data of news commentary are used for lemmas and POS. For the generation step, the table built on the parallel texts of news commentary is augmented with a French dictionary of 280 000 forms. The LM is the largest LM available for French (see section 2.2.2).

4.3 Results and lessons learned

On the news test set of 2008, this system obtains a BLEU score of 20.2, which is worse than our "standard" system (20.9). A similar experiment on the Europarl task proved equally unsuccessful.

Using only models which ignore the surface form of input words yields a poor system. Therefore, including a model based on surface forms, as suggested (Koehn and Hoang, 2007), is also necessary. This indeed improved (+1.6 BLEU for Europarl) over using one single decoding path, but not enough to match our baseline system performance. These results may be explained by the use of automatic tools (POS tagger and lemmatizer) that are not entirely error free, and also, to a lesser extend, by the noise in the test data. We also think that more effort has to be put into the generation step.

Tuning is also a major issue for factored translation models. Dealing with 38 weights is an optimization challenge, which took MERT 129 iterations to converge. The necessary tradeoff between the huge memory requirements of these techniques and computation time is also detrimental to their use.

Although quantitative results were unsatisfactory, it is finally worth mentioning that a manual examination of the output revealed that the explicit usage of gender and number in our models (via POS tags) may actually be helpful when translating to French.

5 Conclusion

In this paper, we presented our statistical MT systems developed for the WMT 08 shared task. As expected, regarding the Europarl condition, our BLEU improvements over the best 2007 results are limited: paying attention to tokenization and caseing issues brought us a small pay-off; rescoring with better language models gave also some reward. The news condition was new, and more challenging: our satisfactory results can be attributed to the use of large, well tuned, language models. In comparison, our experiments with factored models proved disappointing, for reasons that remain to be clarified. On a more general note, we feel that the performance of MT systems for these tasks are somewhat shadowed by normalization issues (tokenization errors, inconsistent use of caseing, typos, etc), making it difficult to clearly analyze our systems' performance.

References

A. Allauzen and H. Bonneau-Maynard. 2008. Training and evaluation of POS taggers on the French multitag corpus. In *Proc. LREC'08, To appear*.

D. Déchelotte. 2007. *Traduction automatique de la parole par méthodes statistiques*. Ph.D. thesis, Univ. Paris XI, December.

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proc. EMNLP-CoNLL*, pages 868–876.

P. Koehn and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proc. of the Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL*, pages 127–133, Edmonton, Canada, May.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, demonstration session*, Prague, Czech Republic.

F.J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. ACL*, pages 295–302.

Franz J. Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL*, Sapporo, Japan.

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