Building a Statistical Machine Translation System for French using the Europarl Corpus

Holger Schwenk

LIMSI-CNRS, bat 508, BP 133
91403 Orsay cedex, FRANCE
schwenk@limsi.fr

Abstract

This paper describes the development of a statistical machine translation system based on the Moses decoder for the 2007 WMT shared tasks. Several different translation strategies were explored. We also use a statistical language model that is based on a continuous representation of the words in the vocabulary. By these means we expect to take better advantage of the limited amount of training data. Finally, we have investigated the usefulness of a second reference translation of the development data.

1 Introduction

This paper describes the development of a statistical machine translation system based on the Moses decoder (Koehn et al., 2007) for the 2007 WMT shared tasks. Due to time constraints, we only considered the translation between French and English. A system with a similar architecture was successfully applied to the translation between Spanish and English in the framework of the 2007 TC-STAR evaluation.[1] For the 2007 WMT shared task a recipe is provided to build a baseline translation system using the Moses decoder. Our system differs in several aspects from this base-line: 1) the training data is not lower-cased; 2) Giza alignments are calculated on sentences of up to 90 words; 3) a two pass-decoding was used; and 4) a so called continuous space language model is used in order to take better advantage of the limited amount of training data.

This architecture is motivated and detailed in the following sections.

2 Architecture of the system

The goal of statistical machine translation (SMT) is to produce a target sentence $\mathbf{e}$ from a source sentence $\mathbf{f}$. 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:

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

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 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 constructed as follows. First, Giza++ is used to perform word alignments in both directions. Second, phrases and lexical reorderings are extracted using the default settings of the Moses SMT toolkit. A target LM is then constructed as detailed in section 2.1. The translation itself is performed in two passes: first, Moses in run and a 1000-best list is generated for each sentence. When generating n-best lists it may happen that the same target sentence is generated multiple times, for instance using different segmentations of the source sentence

[1] A paper on this work is submitted to MT Sumit 2007.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 189–192,
Prague, June 2007. ©2007 Association for Computational Linguistics

or a different set of phrases. We enforced all the hypothesis in an n-best list to be lexically different since our purpose was to rescore them with a LM. The parameters of Moses are tuned on devtest2006 for the Europarl task and nc-dev2007 for the news commentary task, using the cmert tool.

These 1000-best lists are then rescored with different language models, either using a longer context or performing the probability estimation in the continuous space. After rescoring, the weights of the feature functions are optimized again using the numerical optimization toolkit Condor (Berghen and Bersini, 2005). Note that this step operates only on the 1000-best lists, no re-decoding is performed. In general, this results in an increased weight for the LM. Comparative results are provided in the result section whether it seems to be better to use higher order language models already during decoding, or to generate first rich n-best lists and to use the improved LMs during rescoring.

2.1 Language modeling

The monolingual part of the Europarl (38.3M English and 43.1 French words) and the news commentary corpus (1.8M/1.2M words) were used. Separate LMs were build on each data source and then linearly interpolated, optimizing the coefficients with an EM procedure. This usually gives better results than building an LM on the pooled data. Note that we build two sets of LMs: a first set tuned on devtest2006, and a second one on nc-dev2007. It is not surprising to see that the interpolation coefficients differ significantly: 0.97/0.03 for devtest2006 and 0.42/0.58 for nc-dev2007. The perplexities of the interpolated LMs are given in Table 1.

2.2 Continuous space language model

Overall, there are roughly 40 million words of texts available to train the target language models. This is a quite limited amount in comparison to tasks like the NIST machine translation evaluations for which several billion words of newspaper texts are available. Therefore, new techniques must be deployed to take the best advantage of the limited resources.

Here, we propose to use the so-called continuous space LM. The basic idea of this approach is to project the word indices onto a continuous space and to use a probability estimator operating on this space (Bengio et al., 2003). Since the resulting probability functions are smooth functions of the word representation, better generalization to unknown n-grams can be expected. A neural network can be used to simultaneously learn the projection of the words onto the continuous space and to estimate the n-gram probabilities. This is still a n-gram approach, but the LM probabilities are ”interpolated” for any possible context of length n-1 instead of backing-off to shorter contexts.

	French		English	
	Eparl	News	Eparl	News
Back-off LM:				
3-gram	47.0	91.6	57.2	160.1
4-gram	41.5	85.2	51.6	152.4
Continuous space LM:				
4-gram	35.8	73.9	44.5	133.4
5-gram	33.9	71.2	-	-
6-gram	33.1	70.1	41.2	127.0

Table 1: Perplexities on devtest2006 (Europarl) and nc-dev2007 (news commentary) for various LMs.

This approach was successfully used in large vocabulary continuous speech recognition (Schwenk, 2007) and in a phrase-based system for a small task (Schwenk et al., 2006). Here, it is the first time applied in conjunction with a lexicalized reordering model. A 4-gram continuous space LM achieves a perplexity reduction of about 13% relative with respect to a 4-gram back-off LM (see Table 1). Additional improvements can be obtained by using a longer context. Note that this is difficult for back-off LMs due to insufficient training data.

3 Experimental Evaluation

The system was trained on the Europarl parallel texts only (approx. 1.3M words). The news commentary parallel texts were not used. We applied the tokenization proposed by the Moses SMT toolkit and the case was preserved. While case sensitivity may hurt the alignment process, we believe that true case is beneficial for language modeling, in particular in future versions of our system in which we plan to use POS information. Experiences with alternative tokenizations are undergoing.

The parameters of the system were tuned on

	DevTest2006		Test2006	
Decode:	3-gram	4-gram	3-gram	4-gram
Back-off LM:				
decode	30.88	-	30.82	-
4-gram	31.65	31.43	31.35	30.86
Continuous space LM:				
4-gram	31.96	31.75	32.03	31.59
5-gram	31.97	31.86	31.90	31.50
6-gram	**32.00**	31.93	**31.89**	31.64
Lex. diff.	904.2	797.6	900.6	795.8
Oracle	37.82	37.64	-	-

Table 2: Comparison of different translation strategies (BLEU scores for English to French): 3- or 4-gram decoding (columns) and n-best list rescoring with various language models (lines).

	DevTest2006		Test2006	
Decode:	3-gram	4-gram	3-gram	4-gram
Back-off LM:				
decode	32.21	-	31.50	-
4-gram	32.46	32.34	32.07	32.12
Continuous space LM:				
4-gram	32.87	32.90	30.51	32.47
6-gram	32.85	**32.98**	32.46	**32.50**
Lex. diff.	791.3	822.7	802.5	827.8
Oracle	38.80	39.69	-	-

Table 3: Comparison of different translation strategies (BLEU scores for French to English).

devtest2006 and nc-dev2007 respectively. The generalization performance was estimated on the test2006 and nc-devtest2007 corpora respectively.

3.1 Comparison of decoding strategies

Two different decoding strategies were compared in order to find out whether it is necessary to already use higher-order LMs during decoding or whether the incorporation of this knowledge can be postponed to the n-best list rescoring. Tri- or 4-gram back-off language models were used during decoding. In both cases the generated n-best lists were rescored with higher order back-off or the continuous space language model. A beam of 0.6 was used in all our experiments.

The oracle BLEU scores of the generated n-best lists were estimated by rescoring the n-best lists with a cheating LM trained on the development data. We also provide the average number of lexically different hypothesis in the n-best lists. The results are summarized in Table 2 and 3. The numbers in bold indicate the systems that were used in the evaluation.

These results are somehow contradictory : while running Moses with a trigram LM seems to be better when translating from English to French, a 4-gram LM achieves better results when translating to English. An analysis after the evaluation seems to indicate that the pruning was too aggressive for a 4-gram LM, at least for a morphologically rich language like French. Using a beam of 0.4 and a faster implementation of lexical reordering in the Moses decoder, it is apparently better to use a 4-gram LM during decoding. The oracle scores of the n-best lists and the average number of lexically different hypothesis seem to correlate well with the BLEU scores: in all cases it is better to use the system that produced n-best lists with more variety and a higher oracle BLEU score.

The continuous space language model achieved improvements in the BLEU by about 0.4 on the development data. It is interesting to note that this approach showed a very good generalization behavior: the improvements obtained on the test data are as good or even exceed those observed on the Dev data.

3.2 Multiple reference translations

Only one reference translation is provided for all tasks in the WMT’07 evaluation. This may be problematic since systems that do not use the official jargon or different word order may get “incorrectly” a low BLEU score. We have also noticed that the reference translations are not always real translations of the input, but they rely on document wide context information. Therefore, we have produced a second set of sentence based reference translations.[2]

The improvements brought by the continuous space LM are much higher using the new reference translations. Using both reference translations together leads to an important increase of the BLEU score and confirms the improvements obtained by the continuous space LM. These results are in line

[2] The second reference translations can be downloaded from `http://instar.limsi.fr/en/data.html`

Ref. transl.:	official	addtl.	both	retuned
Back-off	31.64	32.91	47.62	47.95
CSLM	32.00	33.81	48.66	49.02

Table 4: Impact of additional human reference translations (devtest2006, English to French)

with our experiences when translating from English to Spanish in the framework of the TC-STAR project (gain of about 1 point BLEU). The BLEU scores can be further improved by rerunning the whole tuning process using two reference translations (last column of Table 4).

Second reference translations for the test data are not yet available. Therefore the devtest data was split into two parts: the back-off and the CSLM achieve BLEU scores of 47.98 and 48.66 respectively on the first half used for tuning, and of 47.95 and 49.02 on the second half used for testing.

3.3 Adaptation to the news commentary task

We only performed a limited domain adaptation: the LMs and the coefficients of the log-linear combination of the feature functions were optimized on nc-dev2007. We had no time to add the news commentary parallel texts which may result in missing translations for some news specific words. The BLEU scores on the development and development test data are summarized in Table 5. A trigram was used to generate 1000-best lists that were then rescored with various language models.

Language modeling seems to be difficult when translating from English to French: the use of a 4-gram has only a minor impact. The continuous space LM achieves an improvement of 0.3 on nc-dev and 0.5 BLEU on nc-devtest. There is no benefit for using longer span LMs. The BLEU score is even 0.5 worse on nc-devtest due to a brevity penalty of 0.95. The continuous space LM also achieves interesting improvements in the BLEU score when translating from French to English.

	English/French		French/English	
	dev	devtest	dev	devtest
Back-off LM:				
decode	27.11	25.31	27.57	26.21
4-gram	27.35	25.53	27.56	26.55
Continuous space LM:				
4-gram	**27.63**	**26.01**	28.25	26.87
6-gram	27.60	25.64	**28.38**	**27.26**

Table 5: BLEU scores for news commentary task.

4 Acknowledgments

This work has been partially funded by the European Union under the integrated project TC-STAR and by the French Government under the project INSTAR (ANR JCJC06_143038). The author would like to recognize the contributions of A. Allauzen for his help with the creation of the second reference translations.

References

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(2):1137–1155.

Frank Vanden Berghen and Hugues Bersini. 2005. CONDOR, a new parallel, constrained extension of powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, 181:157–175.

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 *Proceedings of ACL, demonstation session*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *ACL*, pages 295–302.

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

Holger Schwenk, Marta R. Costa-jussà, and José A. R. Fonollosa. 2006. Continuous space language models for the IWSLT 2006 task. In *IWSLT*, pages 166–173, November.

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