First Steps towards a general purpose French/English Statistical Machine Translation System

Holger Schwenk
LIUM, University of Le Mans
72085 Le Mans cedex 9
FRANCE
schwenk@lium.univ-lemans.fr

Jean-Baptiste Fouet **Jean Senellart**
SYSTRAN SA
92044 Paris La Défense cedex
FRANCE
fouet,senellart@systran.fr

Abstract

This paper describes an initial version of a general purpose French/English statistical machine translation system. The main features of this system are the open-source Moses decoder, the integration of a bilingual dictionary and a continuous space target language model. We analyze the performance of this system on the test data of the WMT'08 evaluation.

1 Introduction

Statistical machine translation (SMT) is today considered as a serious alternative to rule-based machine translation (RBMT). While RBMT systems rely on rules and linguistic resources built for that purpose, SMT systems can be developed without the need of any language knowledge and are only based on bilingual sentence-aligned and large monolingual data. However, while the monolingual data is usually available in large amounts, bilingual texts are a sparse resource for most of the language pairs. The largest SMT systems are currently build for the translation of Mandarin and Arabic to English, using more than 170M words of bitexts that are easily available from the LDC. Recent human evaluations of these systems seem to indicate that they have reached a level of performance allowing a human being to understand the automatic translations and to answer complicated questions on its content (Jones, 2008).

In a joint project between the University of Le Mans and the company SYSTRAN, we try to build similar general purpose SMT systems for European languages. In the final version, these systems will not only be trained on all available mono- and bilingual data, but also will include additional resources from SYSTRAN like high quality dictionaries, named entity transliteration and rule-based translation of expressions like numbers and dates. Our ultimate goal is to combine the power of data-driven approaches and the concentrated knowledge present in RBMT resources. In this paper, we describe an initial version of an French/English system and analyze its performance on the test corpora of the WMT'08 workshop.

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:

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

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 based on the Moses SMT toolkit (Koehn et al., 2007) and constructed as follows.

Proceedings of the Third Workshop on Statistical Machine Translation, pages 119–122,
Columbus, June 2008.

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 4-gram target LM is then constructed as detailed in section 2.2. The translation itself is performed in two passes: first, Moses is run and a 1000-best list is generated for each sentence. The parameters of Moses are tuned on devtest2006 for the Europarl task and nc-devtest2007 for the news task, using the cmert tool. These 1000-best lists are then rescored with a continuous space 5-gram LM and 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. This basic architecture of the system is identical to the one used in the 2007 WMT evaluation(Schwenk, 2007a).

2.1 Translation model

In the frame work of the 2008 WMT shared task, two parallel corpora were provided: the Europarl corpus (about 33M words) and the news-commentary corpus (about 1.2M words). It is known that the minutes of the debates of the European parliament use a particular jargon and these texts alone do not seem to be the appropriate to build a French/English SMT system for other texts. The more general news-commentary corpus is unfortunately rather small in size. Therefore, with the goal to build a general purpose system, we investigated whether more bilingual resources are available. Two corpora were identified: the proceedings of the Canadian parliament, also known as Hansard corpus (about 61M words), and data from the United nations (105M French and 89M English words). In the current version of our system only the Hansard bitexts are used.

In addition to these human generated bitexts, we investigated whether the translations of a high quality bilingual dictionary could be integrated into a SMT system. SYSTRAN provided this resource with more than 200 thousand entries, different forms of a verb or genres of an noun or adjective being counted as one entry. It is still an open research question how to best integrate a bilingual dictionary into a SMT system. At least two possibilities come to mind: add the entries directly to the phrase table or add the words and their translations to the bitexts. With the first solution one can be sure that the entries are added like there are and that they won't suffer any deformation due to imperfect alignment of multi-word expressions. However, it is not obvious how to obtain the phrase translation and lexical probabilities for each new phrase. The second solution has the potential advantage that the dictionary words could improve the alignments of these words when they also appear in the other bitexts. The calculation of the various scores of the phrase table is simplified too, since we can use the standard phrase extraction procedure. However, one has to be aware that all the translations that appear only in the dictionary will be equally likely which certainly does not correspond to the reality. In future work will try to improve these estimates using monolingual data.

For now, we used about ten thousand verbs and hundred thousand nouns from the dictionary. For each verb, we generated all the conjugations in the past, present, future and conditional tense; and for each noun the singular and plural form were generated. In total this resulted in 512k "new sentences" in the bitexts.

2.2 Language model

In comparison to bilingual texts which are needed for the translation model, it is much easier to find large quantities of monolingual data, in English as well as in French. In this work, the following resources were used for the language model:

- the monolingual parts of the Europarl, Hansard, UN and the news commentary corpus,
- the Gigaword corpus in French and English as provided by LDC (770M and 3261M words respectively),
- about 33M words of newspaper texts crawled from the WEB (French only)

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. Note that we build two sets of LMs: a first set tuned on devtest2006, and a second one on nc-devtest2007. The perplexities of these LMs are

	French		English	
Data	Eparl	News	Eparl	News
Back-off 4-gram LM:				
Eparl+news	52.6	184.0	42.0	105.8
All	50.0	136.1	39.7	85.4
Continuous space 5-gram LM:				
All	42.0	118.9	34.1	75.0

Table 1: Perplexities on devtest2006 (Europarl) and nc-devtest2007 (news commentary) for various LMs.

given in Table 1. We were not able to obtain significantly better results with 5-gram back-off LMs.

It can be clearly seen that the additional LM data, despite its considerable size, achieves only a small decrease in perplexity for the Europarl data. This task is so particular that other out-of-domain data does not seem to be very useful. The system optimized on the more general news-commentary task, however, seems to benefit from the additional monolingual resources. Note however, that the test data newstest2008 is not of the same type and we may have a mismatch between development and test data.

We also used a so-called continuous space language model (CSLM). 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. This approach was successfully used in large vocabulary continuous speech recognition (Schwenk, 2007b) and in a phrase-based SMT systems (Schwenk et al., 2006; Déchelotte et al., 2007). Here, it is the first time trained on large amounts of data, more than 3G words for the English LM. This approach achieves an average perplexity reduction of almost 14% relative (see Table 1).

3 Experimental Evaluation

The shared evaluation task of the third workshop on statistical machine translation features two different test sets: test2008 and newstest2008. The first one contains data from the European parliament of the same type than the provided training and development data. Therefore good generalization performance can be expected. The second test set, however, is news type data from unknown sources. Scanning some of the sentences after the evaluation seems to indicate that this data is more general than the provided news-commentary training and development data – it contains for instance financial and public health news.

Given the particular jargon of the European parliament, we decided to build two different systems, one rather general system tuned in nc-devtest2007 and an Europarl system tuned on devtest2006. Both systems use the tokenization proposed by the Moses SMT toolkit and the case was preserved in the translation and language model. Therefore, in contrast to the official BLEU scores, we report here case sensitive BLEU scores as calculated by the NIST tool.

3.1 Europarl system

The results of the Europarl system are summarized in Table 2. The translation model was trained on the Europarl and the news-commentary data, augmented by parts of the dictionary. The LM was trained on all the data, but the additional out-of-domain data has probably little impact given the small improvements in perplexity (see Table 1).

	French/English		English/French	
Model	2007	2008	2007	2008
baseline	32.64	32.61	31.15	31.80
base+CSLM	32.98	33.08	31.63	32.37
base+dict	32.69	32.75	30.97	31.59
+CSLM	33.11	33.13	31.54	32.34

Table 2: Case sensitive BLEU scores for the Europarl system (test data)

When translating from French to English the CSLM achieves a improvement of about 0.4 points BLEU. Adding the dictionary had no significant impact, probably due to the jargon of the parliament proceedings. For the opposite translation direction, the dictionary even seems to worsen the performance. One reason for this observation could be the fact that the dictionary adds many French translations for one English word. These translation are not correctly weighted and we have to rely completely on the target LM to select the correct one. This may explain the large improvement achieved by the CSLM in this case (+0.75 BLEU).

3.2 News system

The results of the more generic news system are summarized in Table 3. The translation model was trained on the news-commentary, Europarl and Hansard bitexts as well as parts of the dictionary. The LM was again trained on all data.

	French/English		English/French	
Model/bitexts	2007	2008	2007	2008
news	29.31	17.98	28.60	17.51
news+dict	30.09	18.78	28.92	18.01
news+eparl	30.53	20.39	28.55	19.70
+dict	30.94	20.63	28.46	19.96
+Hansard	31.48	21.10	28.97	20.21
+CSLM	31.98	21.02	29.64	20.51

Table 3: Case sensitive BLEU scores of the news system (nc-test2007 and newstest2008)

First of all, we realize that the BLEU scores on the out-of-domain generic 2008 news data are much lower than on the nc-test2007 data. Adding more than 60M words of the Hansard bitexts gives an improvement of the BLEU score of about 0.5 for most of the test sets and translation directions. The dictionary is very interesting when only a limited amount of resources is available – a gain of up to 0.8 BLEU when only the news-commentary bitexts are used – but still useful when more data is available. As far as we know, this is the first time that adding a dictionary improved the translation quality of a very strong baseline. In previous works, results were only reported in a setting with limited resources (Vogel et al., 2003; Popović and Ney, 2006). However, we believe that he integration of the dictionary is not yet optimal, in particular with respect to the estimation of the translation probabilities. The only surprising result is the bad performance of the CSLM on the newstest2008 data for the translation from French to English. We are currently investigating this.

This work has been partially funded by the French Government under the project INSTAR (ANR JCJC06_143038).

References

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *JMLR*, 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.

Daniel Déchelotte, Holger Schwenk, Hélène Bonneau-Maynard, Alexandre Allauzen, and Gilles Adda. 2007. A state-of-the-art statistical machine translation system based on moses. In *MT Summit*, pages 127–133.

D. Jones. 2008. DLPT* MT comprehension test results, Oral presentation at the 2008 Nist MT Evaluation workshop, March 27.

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

Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, demonstration 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.

Maja Popović and Hermann Ney. 2006. Statistical machine translation with a small amount of bilingual training data. In *LREC workshop on Minority Language*, pages 25–29.

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. 2007a. Building a statistical machine translation system for French using the Europarl corpus. In *Second Workshop on SMT*, pages 189–192.

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

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *ICSLP*, pages II: 901–904.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical machine translation system. In *MT Summit*, pages 402–409.