SMT and SPE Machine Translation Systems for WMT'09

Holger Schwenk and **Sadaf Abdul-Rauf** and **Loïc Barrault**
LIUM, University of Le Mans
72085 Le Mans cedex 9, FRANCE
schwenk,abdul,barrault@lium.univ-lemans.fr

Jean Senellart
SYSTRAN SA
92044 Paris La Défense cedex, FRANCE
senellart@systran.fr

Abstract

This paper describes the development of several machine translation systems for the 2009 WMT shared task evaluation. We only consider the translation between French and English. We describe a statistical system based on the Moses decoder and a statistical post-editing system using SYSTRAN's rule-based system. We also investigated techniques to automatically extract additional bilingual texts from comparable corpora.

1 Introduction

This paper describes the machine translation systems developed by the Computer Science laboratory at the University of Le Mans (LIUM) for the 2009 WMT shared task evaluation. This work was performed in cooperation with the company SYSTRAN. We only consider the translation between French and English (in both directions). The main differences to the previous year's system (Schwenk et al., 2008) are as follows: better usage of SYSTRAN's bilingual dictionary in the statistical system, less bilingual training data, additional language model training data (*news-train08* as distributed by the organizers), usage of comparable corpora to improve the translation model, and development of a statistical post-editing system (SPE). These different components are described in the following.

2 Used Resources

In the frame work of the 2009 WMT shared translation task many resources were made available. The following sections describe how they were used to train the translation and language models of the systems.

2.1 Bilingual data

The latest version of the French/English Europarl and news-commentary corpus were used. We realized that the first corpus contains parts with foreign languages. About 1200 such lines were excluded.[1] Additional bilingual corpora were available, namely the Canadian Hansard corpus (about 68M English words) and an UN corpus (about 198M English words). In several initial experiments, we found no evidence that adding this data improves the overall system and they were not used in the final system, in order to keep the phrase-table small. We also performed experiments with the provided so-called bilingual French/English Gigaword corpus (575M English words in release 3). Again, we were not able to achieve any improvement by adding this data to the training material of the translation model. These findings are somehow surprising since it was eventually believed by the community that adding large amounts of bitexts should improve the translation model, as it is usually observed for the language model (Brants et al., 2007).

In addition to these human generated bitexts, we also integrated a high quality bilingual dictionary from SYSTRAN. The entries of the dictionary were directly added to the bitexts. This technique has the potential advantage that the dictionary words could improve the alignments of these words when they also appear in the other bitexts. However, it is not guaranteed that multi-word expressions will be correctly aligned by GIZA++ and that only meaningful translations will actually appear in the phrase-table. A typical example is *fire engine – camion de pompiers*, for which the individual constituent words are not good translations of each other. The use of a dictionary to improve an SMT system was also investigated by

[1] Lines 580934–581316 and 599839–600662.

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 130–134,
Athens, Greece, 30 March – 31 March 2009. ©2009 Association for Computational Linguistics

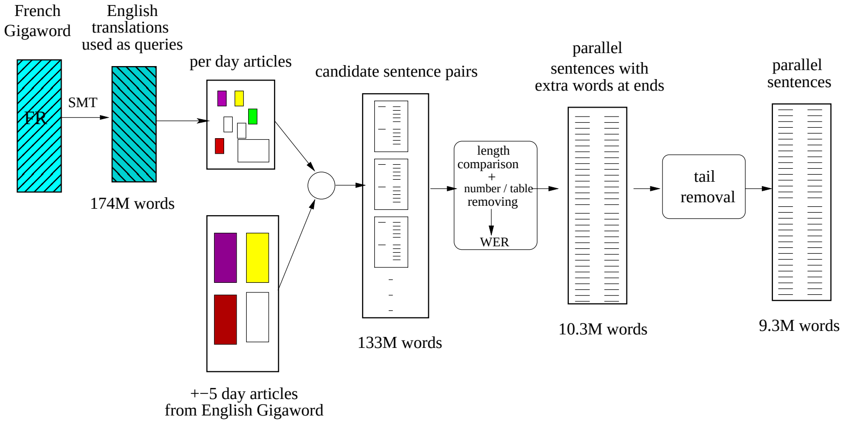
French Gigaword
English translations used as queries
FR
SMT
174M words
per day articles
+−5 day articles from English Gigaword
candidate sentence pairs
133M words
length comparison + number / table removing
WER
parallel sentences with extra words at ends
10.3M words
tail removal
parallel sentences
9.3M words

Figure 1: Architecture of the parallel sentence extraction system (Rauf and Schwenk, 2009).

(Brown et al., 1993).

In comparison to our previous work (Schwenk et al., 2008), we also included all verbs in the French *subjonctif* and *passé simple* tense. In fact, those tenses seem to be frequently used in news material. In total about 10,000 verbs, 1,500 adjectives/adverbs and more than 100,000 noun forms were added.

2.2 Use of Comparable corpora

Available human translated bitexts such as the UN and the Hansard corpus seem to be out-of domain for this task, as mentioned above. Therefore, we investigated a new method to automatically extract and align parallel sentences from comparable in-domain corpora. In this work we used the AFP news texts since there are available in the French and English LDC Gigaword corpora.

The general architecture of our parallel sentence extraction system is shown in figure 1. We first translate 174M words from French into English using an SMT system. These English sentences are then used to search for translations in the English AFP texts of the Gigaword corpus using information retrieval techniques. The Lemur toolkit (Ogilvie and Callan, 2001) was used for this purpose. Search was limited to a window of ±5 days of the date of the French news text. The retrieved candidate sentences were then filtered using the word error rate with respect to the automatic translations. In this study, sentences with an error rate below 32% were kept. Sentences with a large length difference (French versus English) or containing a large fraction of numbers were also discarded. By these means, about 9M words of additional bitexts were obtained. An improved version of this algorithm using TER instead of the word error rate is described in detail in (Rauf and Schwenk, 2009).

2.3 Monolingual data

The French and English target language models were trained on all provided monolingual data. We realized that the *news-train08* corpora contained some foreign texts, in particular in German. We tried to filter those lines using simple regular expressions. We also discarded lines with a large fraction of numerical expressions. In addition, LDC's Gigaword collection, the Hansard corpus and the UN corpus were used for both languages. Finally, about 30M words crawled from the WEB were used for the French LM. All this data predated the evaluation period.

2.4 Development data

All development was done on *news-dev2009a* and *news-dev2009b* was used as internal test set. The default Moses tokenization was used. All our models are case sensitive and include punctuation. The BLEU scores reported in this paper were calculated with the NIST tool and are case sensitive.

3 Language Modeling

Language modeling plays an important role in SMT systems. 4-gram back-off language models (LM) were used in all our systems. The word list contains all the words of the bitext used to train the translation model and all words that appear at least ten times in the *news-train08* corpus. 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. The perplexities of these LMs

Corpus	# Fr words	Dev09a	Dev09b	Test09
SMT system				
Eparl+NC	46.5M	22.44	22.38	25.60
Eparl+NC+dict	48.5M	22.60	22.55	26.01
Eparl+NC+dict+AFP	57.8M	22.82	**22.63***	26.18
SPE system				
SYSTRAN	-	17.76	18.13	19.98
Eparl+NC	45.5M	22.84	**22.59**$^\#$	25.59
Eparl+NC+AFP	54.4M	22.72	21.96	25.40

Table 1: Case sensitive NIST BLEU scores for the French-English systems. "NC" denotes the news-commentary bitexts, "dict" SYSTRAN's bilingual dictionary and "AFP" the automatically aligned news texts (*=primary, $^\#$=contrastive system)

are given in Table 2. Adding the new *news-train08* monolingual data had an important impact on the quality of the LM, even when the Gigaword data is already included.

Data	French	English
Vocabulary size	407k	299k
Eparl+news	248.8	416.7
+ LDC Gigaword	142.2	194.9
+ Hansard and UN	137.5	187.5
news-train08 alone	165.0	245.9
all	120.6	174.8

Table 2: Perplexities on the development data of various language models.

4 Architecture of the SMT system

The goal of statistical machine translation (SMT) is to produce a target sentence e from a source sentence 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:

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

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 based on the Moses SMT toolkit (Koehn et al., 2007) and constructed as follows. First, word alignments in both directions are calculated. We used a multi-threaded version of the GIZA++ tool (Gao and Vogel, 2008).[2] This speeds up the process and corrects an error of GIZA++ that can appear with rare words. This previously caused problems when adding the entries of the bilingual dictionary to the bitexts.

Phrases and lexical reorderings are extracted using the default settings of the Moses toolkit. The parameters of Moses are tuned on *news-dev2009a*, using the cmert tool. The basic architecture of the system is identical to the one used in the 2008 WMT evaluation (Schwenk et al., 2008), but we did not use two pass decoding and n-best list rescoring with a continuous space language model.

The results of the SMT systems are summarized in the upper part of Table 1 and 3. The dictionary and the additional automatically produced AFP bitexts achieved small improvements when translating from French to English. In the opposite translation direction, the systems that include the additional AFP texts exhibit a bad generalisation behavior. We provide also the performance of the different systems on the official test set, calculated after the evaluation. In most of the cases, the observed improvements carry over on the test set.

5 Architecture of the SPE system

During the last years statistical post-editing systems have shown to achieve very competitive performance (Simard et al., 2007; Dugast et al., 2007). The main idea of this techniques is to use

[2] The source is available at `http://www.cs.cmu.edu/~qing/`

Corpus	# En words	Dev09a	Dev09b	Test09
SMT system				
Eparl+NC	41.6M	21.89	21.78	23.80
Eparl+NC+dict	44.0M	22.28	**22.35**$^{\#}$	24.13
Eparl+NC+dict+AFP	51.7M	22.21	21.43	23.88
SPE system				
SYSTRAN	-	18.68	18.84	20.29
Eparl+NC	44.2M	23.03	23.15	24.36
Eparl+NC+AFP	53.3M	22.95	**23.15***	24.62

Table 3: Case sensitive NIST BLEU scores for the English-French systems. "NC" denotes the news-commentary bitexts, "dict" denotes SYSTRAN's bilingual dictionary and "AFP" the automatically aligned news texts (*=primary, $^{\#}$=contrastive system)

an SMT system to correct the errors of a rule-based translation system. In this work, SYSTRAN server version 6, followed by an SMT system based on Moses were used. The post-editing systems uses exactly the same language models than the above described stand-alone SMT systems. The translation model was trained on the Europarl, the news-commentary and the extracted AFP bitexts. The results of these SPE systems are summarized in the lower part of Table 1 and 3. SYSTRAN's rule-based system alone already achieves remarkable BLEU scores although it was not optimized or adapted to this task. This could be significantly improved using statistical post-editing. The additional AFP texts were not useful when translating form French to English, but helped to improve the generalisation behavior for the English/French systems.

When translating from English to French (Table 3), the SPE system is clearly better than the carefully optimized SMT system. Consequently, it was submitted as primary system and the SMT system as contrastive one.

6 Conclusion and discussion

We described the development of two complementary machine translation systems for the 2009 WMT shared translation task: an SMT and an SPE system. The last one is based on SYSTRAN's rule-based system. Interesting findings of this research include the fact that the SPE system outperforms the SMT system when translating into French. This system has also obtained the best scores in the human evaluation.

With respect to the SMT system, we were not able to improve the translation model by adding large amounts of bitexts, although different sources were available (Canadian Hansard, UN or WEB data). Eventually these corpora are too noisy or out-of-domain. On the other hand, the integration of a high quality bilingual dictionary was helpful, as well as the automatic alignment of news texts from comparable corpora.

Future work will concentrate on the integration of previously successful techniques, in particular continuous space language models and lightly-supervised training (Schwenk, 2008). We also believe that the tokenization could be improved, in particular for the French sources texts. Numbers, dates and other numerical expressions could be translated by a rule-based system.

System combination has recently shown to provide important improvements of translation quality. We are currently working on a combination of the SMT and SPE system. It may be also interesting to add a third (hierarchical) MT system.

7 Acknowledgments

This work has been partially funded by the French Government under the project INSTAR (ANR JCJC06_143038) and the by the Higher Education Commission, Pakistan through the HEC Overseas Scholarship 2005.

References

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *EMNLP*, pages 858–867.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer, and Surya Mohanty. 1993. But dictionaries are data too. In *Proceedings of the workshop on Human Language Technology*, pages 202–205, Princeton, New Jersey.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Second Workshop on SMT*, pages 179–182.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

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.

Paul Ogilvie and Jamie Callan. 2001. Experiments using the Lemur toolkit. In *In Proceedings of the Tenth Text Retrieval Conference (TREC-10)*, pages 103–108.

Sadaf Abdul Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *EACL*, page to be published.

Holger Schwenk, Jean-Baptiste Fouet, and Jean Senellart. 2008. First steps towards a general purpose French/English statistical machine translation system. In *Third Workshop on SMT*, pages 119–122.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *IWSLT*, pages 182–189.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In *Second Workshop on SMT*, pages 203–206.

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