LIUM's Statistical Machine Translation Systems for IWSLT 2009

Holger Schwenk, Loïc Barrault, Yannick Estève and Patrik Lambert

LIUM, University of Le Mans, FRANCE

name.surname@lium.univ-lemans.fr

Abstract

This paper describes the systems developed by the LIUM laboratory for the 2009 IWSLT evaluation. We participated in the Arabic and Chinese to English BTEC tasks. We developed three different systems: a statistical phrase-based system using the Moses toolkit, an Statistical Post-Editing system and a hierarchical phrase-based system based on Joshua. A continuous space language model was deployed to improve the modeling of the target language. These systems are combined by a confusion network based approach.

1. Introduction

This paper describes the systems developed by the LIUM laboratory for the 2009 IWSLT evaluation. We extended our work from last year's IWSLT evaluation in several ways: two new systems were built (an hierarchical phrase-based system based on Joshua and a statistical post-editing system), we started to explore various techniques to perform system combination and we participated the first time in the Chinese/English BTEC track.

The reminder of the paper is structured as follows. In sections 2 to 4 we describe the individual systems. Our approach to system combination is explained in section 5 and the experimental results are summarized in section 6. The paper concludes with a discussion of future research issues.

1.1. Used resources

The organizers of IWSLT provide several BTEC specific corpora that can be used to train and optimize the translation system. The characteristics of these corpora are summarized in Table 1. The translation models were trained on the BTEC corpus and the Dev1, Dev2 and Dev3 corpora. The target language model was trained on the English side of the those corpora. No additional texts were used (*constrained condition*). We report results on Dev6 (development data) and Dev7 (internal test set). All BLEU scores are case-sensitive and include punctuations. For some systems, the Dev6 corpus was added to the training material after optimizing the system and the full system was retrained, keeping all settings unmodified. By these means we hope to lower the OOV rate on the official test set. This idea was already successfully proposed in previous IWSLT evaluations [1].

corpus	#lines	#words Arabic	#chars Chinese	#refs
BTEC train	19972	194k	869k	1
Dev1	506	3703	17.7k	16
Dev2	500	3900	17.8k	16
Dev3	506	3801	19.2k	16
Dev6	489	3612	16.5k	6
Dev7	500	3931	17.4k	16
Eval09	469	3494	15.9k	n/a

Table 1: Characteristics of the provided BTEC data. Dev6 was used to optimize our systems and Dev7 as internal test set.

1.2. Tokenization

The Arabic texts were tokenized using the sentence analysis module of SYSTRAN's rule-based Arabic/English translation software. Sentence analysis represents a large share of the computation in a rule-based system. This process applies first decomposition rules coupled with a word dictionary. For words that are not known in the dictionary, the most likely decomposition is guessed. In general, all possible decompositions of each word are generated and then filtered in the context of the sentence. This steps uses lexical knowledge and a global analysis of the sentences. In a similar way, the Chinese texts were segmented into words using tools from SYSTRAN.

2. SMT System

The statistical phrase-based system is based on the Moses SMT toolkit [2] and constructed as follows. First, Giza++ is used to perform word alignments in both directions. Second, phrases and lexical reorderings are extracted. Both steps use the default settings of the Moses SMT toolkit. A 4-gram back-off target language model (LM) is constructed on all available English data. The translation itself is performed in two passes: first, Moses is run and a 1000-best list is generated for each sentence. 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. The coefficients of these feature functions are tuned on development data using the cmert tool.

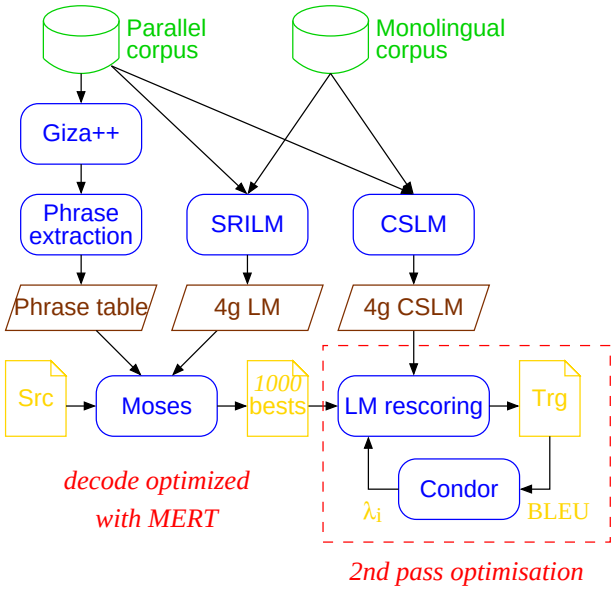

Figure 1: Architecture of the SMT system.

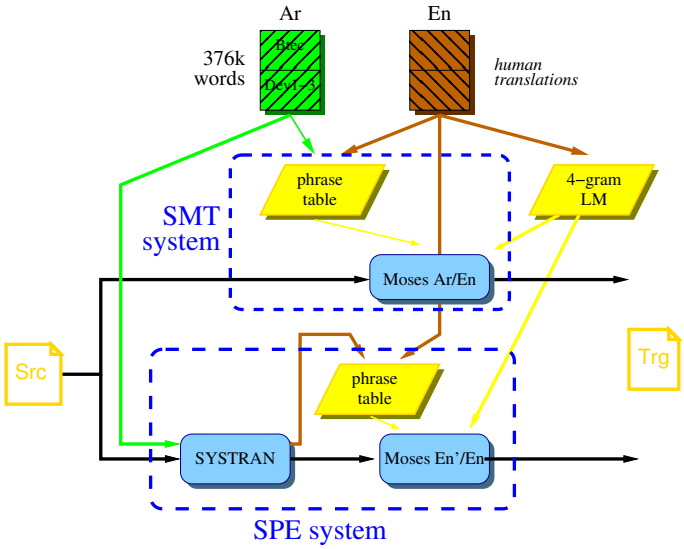

Figure 2: Comparison of SMT and SPE systems.

These 1000-best lists are then rescored with a continuous space 4-gram LM and the weights of the feature functions are again optimized, this time using the open source numerical optimization toolkit Condor [3]. This basic architecture of the system is summarized in Figure 1.

2.1. Continuous space language model

For the BTEC task, there are less than 400k words of in-domain English 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, as in previous work, we applied the so-called continuous space language model. 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 [4]. 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 language model posterior probabilities are "interpolated" for any possible context of length $n-1$ instead of backing-off to shorter contexts. This approach is expected to take better advantage of the limited amount of training data.

Training is performed with the standard back-propagation method using weight decay and a re-sampling algorithm [5]. In our experiments, a reduction of the perplexity of about 17% was obtained.

3. SPE System

In the last years, there is increasing interest in the interaction between rule-based and statistical machine translation. A popular and successful idea is *statistical post editing* (SPE) [6, 7]. The principle idea is to train an SMT system to correct the outputs of a rule-based translation system. This is shown in figure 2. The operation performed by the rule-based translation system could also be seen as a very good tokenization or preprocessing, that actually performs many of the translation steps. Therefore, the task of the SMT system itself is very simplified. Accordingly, we argue that an SMT and SPE system are only two extreme cases of the interaction between tokenization/preprocessing and translation itself. An interesting question is whether both systems can be combined since the SPE system somehow already includes an SMT system. This was investigated in this work.

4. Hierarchical System

Hierarchical phrase-based translation systems [8] are an interesting alternative to the standard phrase-based systems. Recently, the large-scale parsing-based statistical machine translation tool "Joshua" has been made available under a open source license [9]. This decoder is written in Java and is scalable by the use of parallel and distributed computing. In [10], Joshua achieves competitive results in comparison to Moses and we have decided to experiment with Joshua during the IWSLT campaign.

The entire statistical system based on Joshua was built in a similar way as the SMT system. First, the tool "berkeleyAligner" is used to perform alignments instead of Giza++, according to the recommendations in the documentation of the Joshua tool kit. Then, grammar rules are extracted from these alignments. The same 4-gram LM as in the SMT sys-

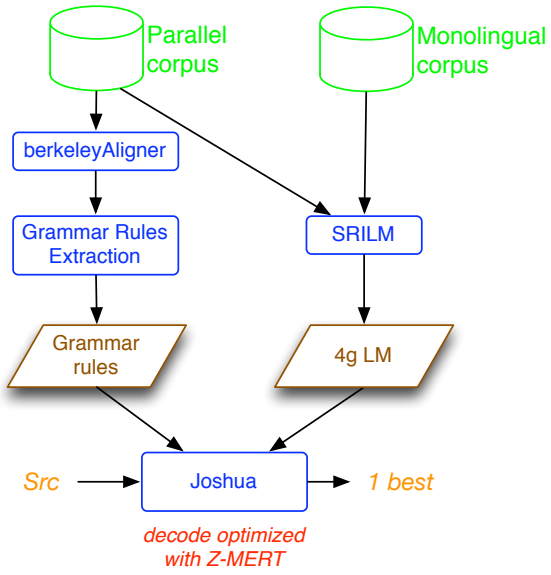

Figure 3: System architecture of the hierarchical system – see text for details.

tem was used with Joshua. Weights were optimized on the development set (Dev6) using the provided Z-MERT procedure. The grammar rules extraction tools and Z-MERT are provided in the Joshua toolkit. Figure 3 summarizes the architecture of the Joshua translation system.

5. System combination

The system combination approach is based on confusion network decoding as described in [11, 12] and shown in Figure 4. The protocol can be decomposed into three steps :

1. 1-best hypotheses from all M systems are aligned and confusion networks are built.
2. All confusion networks are connected into a single lattice.
3. A 4-gram language model is used to decode the resulting lattice and the best hypothesis is generated.

5.1. Hypotheses alignment and confusion network generation

For each segment, the best hypotheses of $M-1$ systems are aligned against the last one used as backbone. The alignment is done with the TER tool [13], without any tuning performed at this step (default edit costs are used). M confusion networks are generated in this way. Then all the confusion networks are connected into a single lattice by adding a first and last node. The probability of the first arcs must reflect how well such system provides a well structured hypothesis (good order). In our experiments, no tuning was done at this step, and we chose equal prior probabilities for all systems.

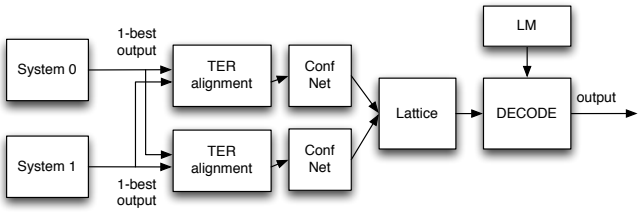

Figure 4: MT system combination.

A preliminary version of our system combination tools were used during this evaluation period and only two systems could be combined. Creating a confusion network based on more than one alignment is not obvious and some decisions have to taken in account to efficiently merge the alignments. When combining two systems, the confusion networks are built directly from the result of the alignment (which is trivial in this case). Also, this version does not use a translation score for each word, as provided by the individual translation systems. Instead, we used weights equal to the priors.

5.2. Decoding

The decoder is based on the token pass decoding algorithm. The scores used to evaluate the hypotheses are the following:

- the system score : this replaces the score of the translation model. Until now, the words given by all systems have the same probability which is $\frac{1}{M}$.
- the language model (LM) probability. The 4-gram LM used for the combination is the same than the one used by each single system.

It is obvious that this combination framework is not optimal, but as we can see in the results section, this simple architecture can already achieve improvements when combining only two systems.

6. Experimental Evaluation

The case-sensitive BLEU scores for the various systems are summarized in Table 2. The Moses phrase-based systems achieved the best performance for both language pairs. This is contrast to other studies which report that hierarchical systems outperform phrase-based systems, in particular when translating from Chinese to English. We are currently investigating how to better optimize our hierarchical systems built with Joshua.

Rescoring the n-best lists with the continuous space LM achieved an improvement of 1.2 BLEU on the internal test set for the Arabic/English SMT system, and 0.6 BLEU for the SPE system. Due to time constraints, the continuous space LM was not applied on the hierarchical system. The improvements obtained by the CSLM are generally smaller

Approach:		SMT *Moses*		Hierarchical *Joshua*		SPE *SYSTRAN + Moses*	
Train bitexts	LM	Dev	Test	Dev	Test	Dev	Test
Arabic/English:							
Btec+Dev123	back-off	53.58	53.41	53.05	53.49	50.22	47.55
	CSLM	54.54	54.61	-	-	51.31	**48.13**
Btec+Dev1236	back-off	-	-	n/a	**54.00**	-	
	CSLM	n/a	**54.75**	-	-		
Chinese/English:							
Btec+Dev1-3	back-off	33.30	41.29	28.54	**39.78**	29.32	40.83
	CSLM	33.65	**41.71**	-	-	30.90	**41.23**

Table 2: Comparison of the BLEU scores of all the systems. The systems marked in bold were used for system combination. All systems are tuned on Dev6 and tested on Dev7. CSLM denotes the continuous space language model.

when translating from Chinese to English: 0.4 BLEU for both the SMT and SPE system.

Adding the Dev6 data to the bitexts was only performed for the Arabic/English systems. It yielded an improvement of 0.5 BLEU points for the hierarchical system, but not significant gain for the SMT system.

6.1. System Combination

In all the system combination experiments, the corpus Dev7 was used as development corpus in order to select the two systems to combine. The results are presented in table 3

Systems		Arabic/English	Chinese/English
SMT	back-off	53.41	41.29
	CSLM	**54.75**	**41.71**
SPE	back-off	47.55	40.83
	CSLM	48.13	41.23
Hierarchical	back-off	54.00	39.78
	CSLM	-	-
SMT + SPE	back-off	54.34	39.63
	CSLM	54.40	**42.55**
SMT + Hier.	back-off	55.54	40.30
	CSLM	**55.89**	40.18
SPE + Hier.	back-off	51.62	38.95
	CSLM	54.84	-

Table 3: System combination on Dev7 corpus.

When the initial systems have a different level of performance, then the result of the combination is hardly predictable. For example, when combining SMT and SPE for Zh/En, an improvement of 0.84 BLEU points (respectively to the best single system) is observed even if SPE system alone performs much worse. On the contrary, the combination of SMT and SPE for Ar/En decreases the results.

This is not the case when systems have similar performances. It appears that the combination of the best 2 systems gives the best performance. In the end, systems SMT + Joshua for Ar/En and SMT + SPE for Zh/En were selected for combination.

6.2. Performance on the evaluation data

The best performing system combinations were submitted as primary systems to the IWSLT 2009 evaluation. In addition, the following contrastive runs were submitted for scoring:

- The individual SMT and SPE systems for Arabic/English
- The SMT, SPE and hierarchical systems for Chinese/English

The results provided by the organizers are summarized in table 4. There are several notable differences in comparison to the performances observed on the internal test data. First of all, for Arabic/English system combination did not work very well on the official test set: we only achieve an improvement of 0.5 BLEU with respect to the best individual system. There was a gain of 1.1 BLEU points on the internal test set. This may be explained by the fact the hierarchical system seems to perform badly on the official test data: it is 1.3 BLEU points worse than the SMT system.

Looking at Chinese/English, we observe the opposite effect: system combination works better on the official test set (+1.6 BLEU) in comparison to the internal test set (+0.8 with respect to the best individual system). Again, this may be explained by the performance of the individual systems. It appears in fact the the SPE system achieves better result on the official test data than the SMT system. We are currently investigating the possible reasons for those observations.

7. Conclusion and discussion

This paper described the statistical machine translation systems developed by the LIUM laboratory for the 2009 IWSLT evaluation. We participated in the BTEC Arabic and Chinese/English tasks. For both language pairs, an SMT system based on Moses, an hierarchical system based on Joshua and an SPE system was developed. Initial system combination

	BLEU	meteor	f1	prec	recl	wer	per	ter	gtm	nist
Arabic/English										
primary (SMT+Hier)	**0.5086**	0.7315	**0.7789**	**0.8238**	0.7387	0.3669	0.3295	**30.3340**	0.7460	7.1976
contrastive1 (SMT)	0.5035	**0.7397**	0.7762	0.7981	**0.7554**	**0.3643**	**0.3247**	30.6900	**0.7544**	**7.7605**
contrastive2 (Hier)	0.4906	0.7306	0.7743	0.8084	0.7429	0.3788	0.3391	31.2500	0.7400	7.3100
Chinese/English										
primary (SMT+SPE)	**0.4014**	0.6076	0.6653	**0.7143**	0.6226	**0.4921**	**0.4378**	**41.4800**	**0.6768**	6.1194
contrastive1 (SMT)	0.3604	0.5958	0.6546	0.6955	0.6182	0.5310	0.4586	45.3230	0.6708	6.1984
contrastive2 (SPE)	0.3853	**0.6428**	**0.6788**	0.6809	**0.6767**	0.5035	0.4389	43.3890	0.6743	**6.9109**
contrastive3 (Hier)	0.3189	0.5623	0.6431	0.7140	0.5850	0.5596	0.4861	45.0430	0.6406	4.5253

Table 4: Results on the official 2009 test data

experiments yielded improvements in the BLEU score of up to 1.6 BLEU points.

After the official evaluation period, we added some features to our system combination scheme. In the decoder, a fudge factor has been included in order to weight the probabilities given by the language model and those available in the lattice. Moreover, a null-arc and a length penalty have been added. The probabilities computed in the decoder can now be expressed as follow :

$$log(P_W) = \sum_{n=0}^{Len(W)} [log(P_{ws}(n)) + \alpha P_{lm}(n)] + Len_{pen}(W) + Null_{pen}(W) \quad (1)$$

where $Len(W)$ is the length of the hypothesis, $P_{ws}(n)$ is the score of the n^{th} word, $P_{lm}(n)$ is its LM probability, $Len_{pen}(W)$ is the length penalty of the word sequence and $Null_{pen}(W)$ is the penalty associated with the number of null-arcs crossed to obtain the hypothesis.

Those features have been tuned using the Dev7 corpus for the Arabic-English task. The official test set has been reprocessed with this new setup and a BLEU score of **51.74** was obtained. This is an improvement of 0.88 BLEU points compared to the previous system combination and of 1.39 relatively to the best single system. The next step will be to enable the combination of more than two systems.

7.1. Acknowledgments

This work has been partially funded by the French Government under the project INSTAR (ANR JCJC06 143038) and the European project FP7 EUROMATRIXPLUS. We are very thankful to Jean Senellart and Jean-Baptiste Fouet from the company SYSTRAN S.A. who provided the support for the Arabic and Chinese tokenization.

8. References

[1] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney, "The RWTH statistical machine translation system for the IWSLT 2006 evaluation," in *IWSLT*, 2007, pp. 103–100.

[2] P. Koehn et al., "Moses: Open source toolkit for statistical machine translation," in *ACL, demonstration session*, 2007.

[3] F. V. Berghen and H. Bersini, "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*, vol. 181, pp. 157–175, 2005.

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *JMLR*, vol. 3, no. 2, pp. 1137–1155, 2003.

[5] H. Schwenk, "Continuous space language models," *Computer Speech and Language*, vol. 21, pp. 492–518, 2007.

[6] L. Dugast, J. Senellart, and P. Koehn, "Statistical post-editing on SYSTRAN's rule-based translation system," in *Second Workshop on SMT*, 2007, pp. 179–182.

[7] M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn, "Rule-based translation with statistical phrase-based post-editing," in *Second Workshop on SMT*, 2007, pp. 203–206.

[8] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[9] Z. Li, C. Callison-Burch, S. Khudanpur, and W. Thornton, "Joshua: Open source, parsing-based machine translation," *The Prague Bulletin of Mathematical Linguistics*, no. 91, 2009.

[10] B. Chen, D. Xiong, M. Zhang, A. Aw, and H. Li, "I2r multi-pass machine translation system for iwslt 2008," in *IWSLT*, 2008.

[11] W. Shen, B. Delaney, T. Anderson, and R. Slyh, "The MIT-LL/AFRL IWSLT-2008 MT system," in *IWSLT*, Hawaii, U.S.A, 2008, pp. 69–76.

[12] A.-V. Rosti, S. Matsoukas, and R. Schwartz, "Improved word-level system combination for machine translation," in *ACL*, 2007, pp. 312–319.

[13] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *ACL*, 2006.