LIUM's Statistical Machine Translation Systems for IWSLT 2010 Talk Task

Anthony Rousseau, Loïc Barrault, Yannick Estève and Paul Delglise

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 2010 IWSLT evaluation. We participated in the English/French Talk Task. We developed a statistical phrase-based system using the Moses toolkit.

1. Introduction

This paper describes the systems developed by the LIUM laboratory for the 2010 IWSLT evaluation. We participated in the new task of this year evaluation, *i.e.* the Talk Task, which consists in translation shows from the TED website.

The remainder of the paper is structured as follows. In sections 2 to **??** we describe the individual systems. Specific strategies for translation in ASR conditions are described in section **??** and the experimental results are summarized in section 4. The paper concludes with a discussion on future research issues.

1.1. Used resources

The organizers of IWSLT provide several 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 [**?**].

1.1.1. Monolingual data

1.1.2. Bitexts

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.

corpus	#lines	#tok English	#tok French
TED train v1.1	84.5k	877k	943k
news-commentary10	84.6k	2M	2.4M
europarl.v5	1.6M	45M	45M
un200x	7.2M	211.7M	240.2M
Gigaword_fr-en	22.5M	662.7M	771.7M
TED dev CRR	1307	12554	12528
TED dev ASR 1Best	259	11334	n/a
TED test CRR	3502	31980	n/a
TED test ASR 1Best	758	28115	n/a

Table 1: Characteristics of the provided data.

2. SMT System

The statistical phrase-based system is based on the Moses SMT toolkit [**?**] 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. 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 [**?**]. This basic architecture of the system is summarized in Figure 1.

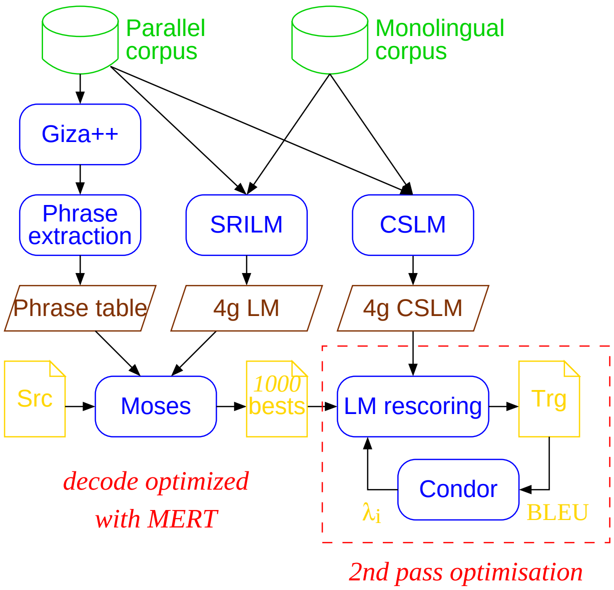

Figure 1: Architecture of the SMT system.

3. System combination

The system combination approach is based on confusion network decoding as described in [?, ?] and shown in Figure 2. 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.

3.1. 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 [?], 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 provide 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.

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 bu the individual translation systems. Instead, we used weights equal to the priors.

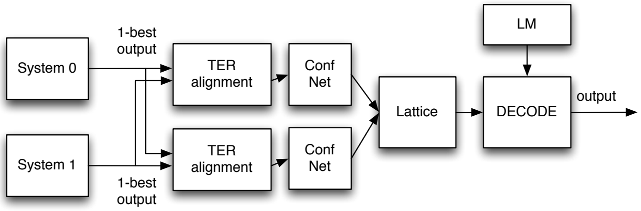

Figure 2: MT system combination.

3.2. 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 replace 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.

4. 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 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.

Approach:		SMT		Hierarchical		SPE	
		Moses		*Joshua*		*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.

4.1. 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 3. 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.

5. 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 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 :

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

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.

5.1. Acknowledgments

	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 3: Results on the official 2009 test data