LIUM's Statistical Machine Translation Systems for IWSLT 2009

Holger Schwenk, Loïc Barrault, Yannick Estève and Patrik Lambert

LIUM, University of Le Mans

Surname.Name@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/English BTEC tasks. We developed three different systems: a statistical phrase-based system using the Moses toolkit, an Statistical Post-Editing (SPE) 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.

INTRODUCTION

New features with respect to last year's system:

- Arabic/English BTEC task and
  first participation in the Chinese/English BTEC track.
- Bitexts and LM resources limited to provided BTEC data
- Classical SMT system based on Moses
- First experiments with Joshua-based hierarchical system
- Development of a statistical post-editing system (SPE)
⇒ First steps in system combination

RESOURCES

Characteristics of the provided BTEC data

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

- Training on Btec + Dev1–3
- Development on Dev6, internal test on Dev7
- For some systems, Dev6 was added to the training material after tuning, keeping all settings unmodified.
- The Arabic texts were tokenized using SYSTRAN's sentence analysis module. It includes a morphological decomposition.
- Chinese characters were segmented using tools from SYSTRAN
- All models are case-sensitive and with punctuations

SYSTEM ARCHITECTURE

SMT system

- Statistical phrase-based system using Moses and own tools
- Two pass approach:
  - Generation of 1000-best lists with 4-gram back-off LM
  - Rescoring of those n-best lists with continuous space LM

SPE System

- Use of an SMT system to correct the errors of a rule-base system
  [Dugast et al, WMT'07, Simard et al, WMT'07]
- Here: SYSTRAN version 6 + Moses
- The LM is identical to the one used in the stand-alone SMT system
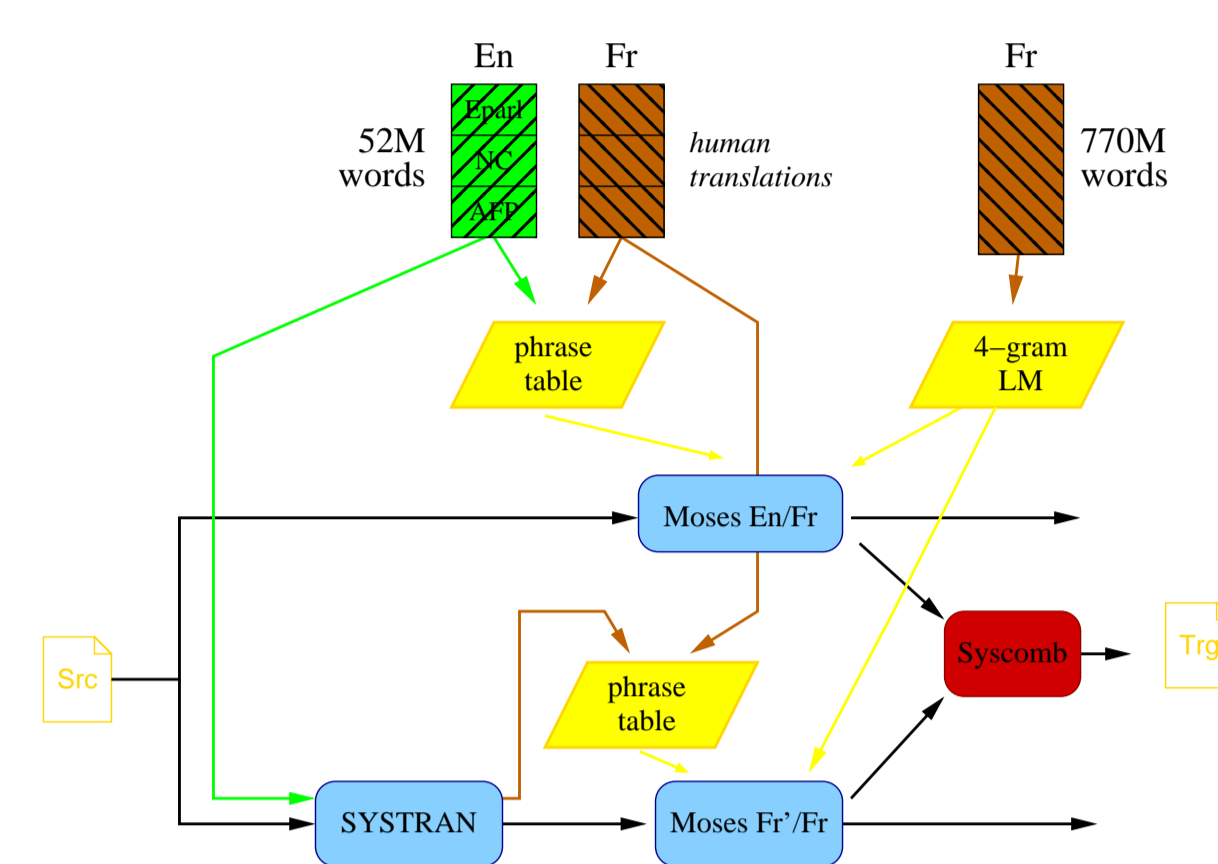
Hierarchical system

- First experiments with the Joshua toolkit from JHU
- Alignment using the *BerkeleyAligner*
- Extraction of grammar rules
- Weight tuning with the provide z-mert tool
- LM identical to the one of the other two systems

SYSTEM COMBINATION

The system combination approach is based on confusion network decoding similar to previous work:

1. 1-best hypotheses from all M systems are aligned using TER and confusion networks are built.
2. All confusion networks are connected into a single lattice. (the initial version of our system did not use weights for each branch)
3. A 4-gram LM is used to decode the resulting lattice and the best hypothesis is generated.
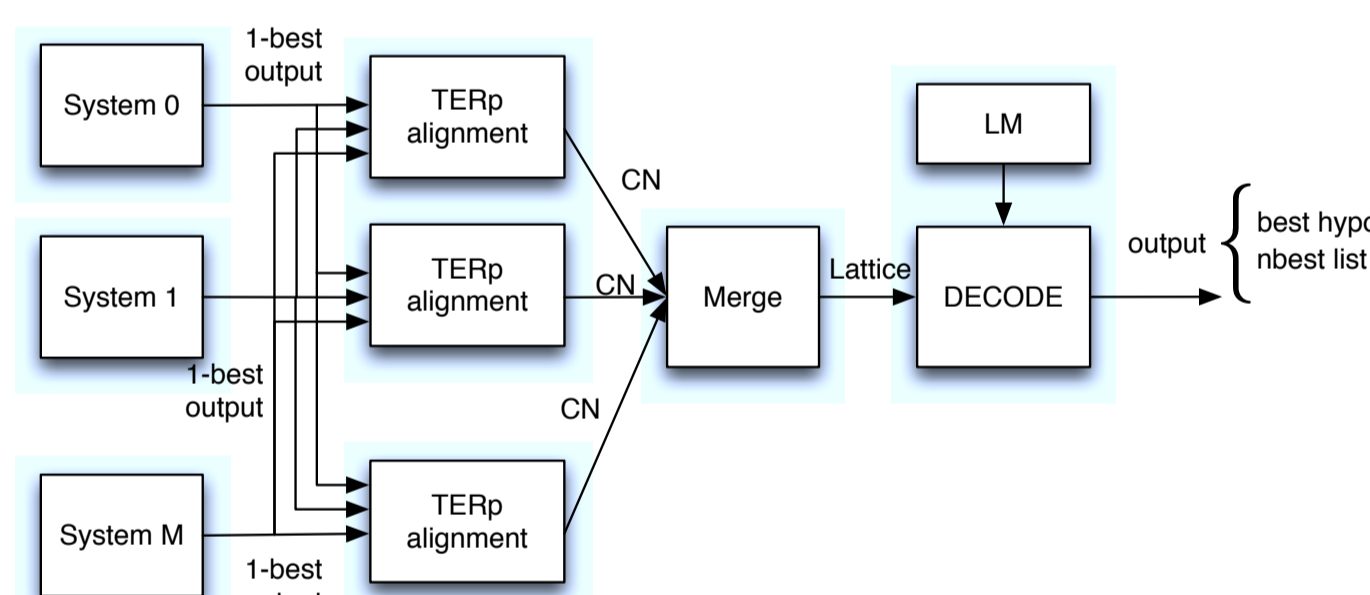
Scores and parameters

- Loïc: we need some details here (1st and final version)

EXPERIMENTAL EVALUATION

Case-sensitive BLEU scores of all the systems

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**	n/a	-		
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**

- The Moses phrase-based systems achieved the best performance for both language pairs
- Joshua gives very competitive results for Ar/En, but less so for Zh/En
- SPE systems is quite interesting for Chinese/English
- CSLM achieved improvements of up to 1.2 BLEU as in previous experiments
- The CSLM was not yet applied on the hierarchical system
- Adding Dev6 to the bitexts helped only the hierarchical system

System combination results on Dev7

- Official evaluation results with the simplified approach

Systems		Arabic/English	Chinese/English
SMT	back-off	53.41	41.29
	CSLM	**54.75**	**41.71**
SPE	back-off	46.13	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	-

- Improvements obtained by system combination:
  +1.14 for Arabic-English and +0.84 for Chinese-English.
- Important to notice that only 2 systems were involved and no tuning was performed at this stage.

Results on the official 2009 test data

Systems	Arabic/English	Chinese/English
SMT CSLM	50.35	36.04
SPE CSLM	-	38.53
Hierarchical	49.06	31.89
SMT CSLM + SPE CSLM	-	40.14
SMT CSLM + Hier.	50.86	-

- SPE achieves very good performance on Zh/En in comparison to Moses alone (less OOVs ?)
- There seem to be some problems with our hierarchical system: results degrade a little on Ar/En and substantially on Zh/En
- Improvements of up to 1.7 BLEU by simple system combination of two systems only

Experiments after the evaluation

- Improved system combination including tuning of weights:

Systems	Arabic/English Dev7	Test09	Chinese/English Dev7	Test09
SMT CSLM + SPE CSLM	-	-	-	-
SMT CSLM + Hier.				
+ manual tuning	57.01	51.74	-	-
+ tuning with Condor	57.27	51.65	-	-

- Improvement of 1.6 BLEU on Ar/En (instead of 0.5 BLEU)
- LOIC: have you done it for Zh/En ??

CONCLUSION AND PERSPECTIVES

- Development of three complementary systems: Moses, Joshua and SPE
- The Moses system is overall still the best one
- But the SPE showed a very good generalization behavior on the eval data.
- Continuous space LM was useful as in previous experiments
- Initial work in system combination

Ongoing work

- Continuous space LM
- Unsupervised training on *news-train08*
- System combination

ACKNOWLEDGMENTS