LIUM's Statistical Machine Translation System for IWSLT 2010

Anthony Rousseau, Loïc Barrault, Paul Deléglise, Yannick Estève

LIUM, University of Le Mans,
72085 Le Mans cedex 9, France
FirstName.LastName@lium.univ-lemans.fr

ABSTRACT

Participation of LIUM to the 2010 IWSLT campaign:

- TALK task (based on TED website talks).
- One system for each text condition.
- Specific strategies for ASR text condition.
- Experiments on handling ASR word lattices.

INTRODUCTION

The new 2010 IWSLT TALK Task:

- English to French task.
- Constrained condition.
- Based on TED talks (http://www.ted.com).
- Wide variety of speakers and native languages.
- Two submissions required:
  - Correct recognition results, *i.e.* ASR reference (CRR),
  - Automatic speech recognition outputs (ASR).

The LIUM's systems emphasize on:

- Adaptation to ASR condition:
  - SMT system trained on ASR-resembling text,
  - case & punctuation treated by a statistical approach.
- Handling ASR lattices:
  - reduction in size,
  - transformation in confusion networks (CNs).
- Rescoring with part-of-speech LM:
  - compute 7-gram POS LM,
  - add a POS score to SMT hypothesis then rescore.

RESOURCES

Bilingual data

Available corpora:

corpus	#lines	#tok English	#tok French
TED v1.1	84.5k	877k	943k
News-Commentary 10	84.6k	2M	2.4M
Europarl v5	1.6M	45M	45M
UN200x	7.2M	211.7M	240.2M
Gigaword release 2	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

Monolingual data

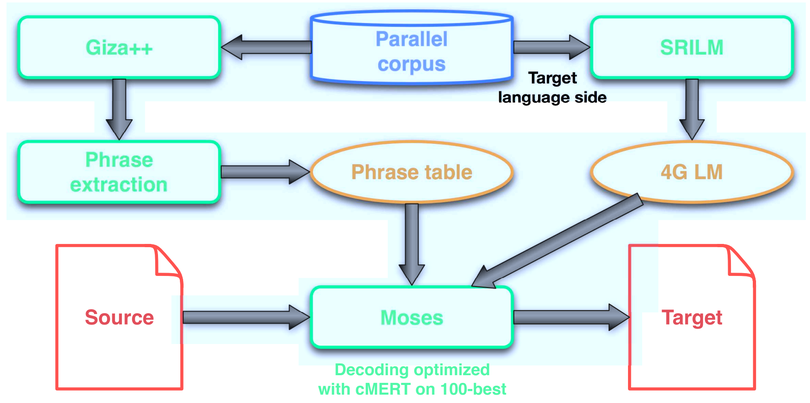
Target LMs trained on French sides of the proposed bitexts.

Data selection and filtering

- Filtering performed with lexical costs of sentence pairs.
- Data selection based on BLEU scores of corpora subsets.

ARCHITECTURE OF THE SYSTEMS

- PBSMT system using Moses (default settings).
- Alignments in both directions with Giza++.
- Fourteen feature functions:
  - phrase and lexical translation probs (both directions),
  - seven features for the lexicalized distortion model,
  - word and phrase penalty,
  - target LM.
- Default Moses tokenisation.
- 4-gram backoff LMs with SRILM:
  - one LM on each corpora then linear interpolation.
- Coefficients optimized with cMERT on 100-best lists.



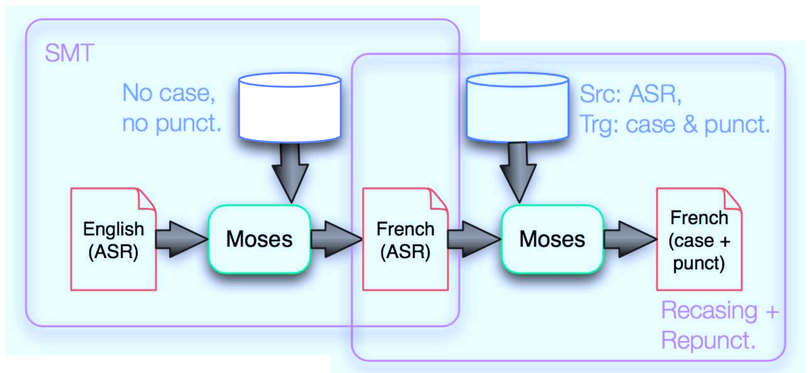
ADAPTATION TO ASR SPECIFICITIES

ASR outputs usually:

- are lowercased,
- contain no punctuation,
- differs from SMT on normalization:
  - written numbers,
  - acronyms,
  - contractions.

Our approach:

- Create a parallel corpus which resembles ASR outputs:
  - suppress all punctuation,
  - lowercase all words (with some exceptions),
  - transform numbers into letters,
  - normalize many contractions and symbols.
- Train a SMT system on this corpus.
- Optimize it on the provided 1-best development corpus.
- Estimate a separate LM with no punctuation nor case.
- Treat the case and punctuation issues:
  - create a new bitext from original and ASR corpora (French),
  - train a new system with it,
  - optimize on the CRR dev corpus (with case & punctuation),
  - decode the translation output in ASR condition.
    ⇒ Necessity to limit the distortion.



⇒ Leads to a cased and punctuated output.

OFFICIAL RESULTS

	WER	dev set		test set	
		CRR condition 1			
-	-	BLEU	TER	BLEU	TER
		26.45	61.02	25.07	57.60
		ASR condition 1			
1-Best	24.8	BLEU	TER	BLEU	TER
		16.82	70.86	15.82	71.15
		ASR condition 3			
		BLEU	TER	BLEU	TER
		18.49	70.01	18.27	70.92

Condition 1: with punctuation and casing.
Condition 3: no punctuation, no case.

HANDLING ASR LATTICES

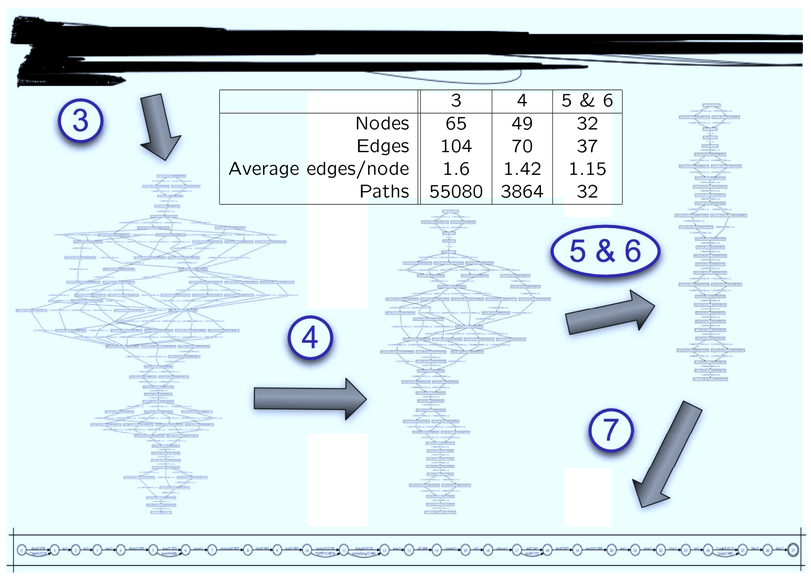
Some information about the lattices building is missing:

- word insertion penalty,
- linguistic weight used (two are provided).

Provided lattices too large to be managed directly by Moses.
⇒ Necessity to reduce their size.

This reduction can be summarized by these steps:

1. Compute link posteriors with *forward-backward* algo.
2. Split some words to normalize the lattice tokenization.
3. Merge identical words located in equivalent temporal area.
4. Prune links with posteriors < .001. Repeat step 3.
5. Prune links with posteriors < .01. Repeat step 3.
6. Remove filler words and ϵ (null transitions).
7. Transform the PLF lattice in confusion network and write both of them.



	3	4	5 & 6
Nodes	65	49	32
Edges	104	70	37
Average edges/node	1.6	1.42	1.15
Paths	55080	3864	32

RESCORING WITH PART-OF-SPEECH LM

- Tag n-best SMT hypotheses and French corpora with *lia_tagg*.
- Compute a 7-gram POS LM on the POS-tagged training data.
- Add a POS LM score to each n-best SMT hypothesis.
- Recompute the global score of each hypothesis with optimized linear coefficients.

	dev set	test set
Best point without POS	19.44	**20.98**
Best point after tuning	**19.79**	20.65

⇒ This approach does not generalize very well.
⇒ Further analysis of tags needed to understand these results.

LATE RESULTS

	WER	ASR condition 3			
		dev set		test set	
PLF	26.4	BLEU	TER	BLEU	TER
		19.44	69.33	**20.98**	66.09
		dev set		test set	
CN	26.1	BLEU	TER	BLEU	TER
		19.39	69.39	-	-
		dev set		test set	
1-Best	24.8	BLEU	TER	BLEU	TER
		19.19	69.45	20.14	66.77

CONCLUSION

- POS LM rescoring needs further investigation.
- Confusion networks weights tuning is not optimal.
- Bigger search space as SMT input leads to improvement:
  - compared to 1-best,
  - even when WER is higher.