The LIGA (LIG/LIA) Machine Translation System for WMT 2011

Marion Potet, Raphaël Rubino, Benjamin Lecouteux, Stéphane Huet, Hervé Blanchon, Laurent Besacier, Fabrice Lefèvre

The LIGA (LIG/LIA) Machine Translation System for WMT 2011

Marion Potet, Raphaël Rubino, Benjamin Lecouteux, Stéphane Huet, Hervé Blanchon, Laurent Besacier and Fabrice Lefèvre

Introduction

- The **LIG and LIA laboratory** have combined their efforts to produce a **joint submission** to WMT 2011 for the **French-English translation task**.

- Each group started by developing its own solution whilst sharing ressources (corpora as provided by the organizers, aligned data, etc.) and acquired knowledge (current parameters, effect of the size of n-grams, etc.) with the other.

→ The final LIGA submission is a **combination of two phrase-based translation systems** with appropriate carefully-tuned setups.

LIA system

- Pre-processing:
  . **truecased** text
  . **Reaccentuation of French words** starting with a capital letter

- **5-gram LM**: *mono-news-c + news-s*

- **Translation model:** Training on 10M sentence pairs selected in *news-c + euro + UN + giga*

Used parallel corpora	Filtering: without	Filtering: with
news-c + euro (1.77 M)	28.1	28.0
news-c + 1.77 M of *UN*	27.2	-
news-c + 1.77 M of *giga*	27.1	-
news-c + 1.77 M with IR	28.2	-
news-c + 3 M with IR	29.1	29.0
news-c + 5 M with IR	28.8	-
***news-c* + 10 M with IR**	**29.3**	29.2
All data	28.9	29.0

Table : BLEU (%) on test10 measured with the LIA system using different training parallel corpora

- **Phrase based translation model**
- Tokenisation of corpora
- Translation model: **Giza++** and **Moses**
- **Decoding: 14 scores** on features function
- **Language model** : n-gram based, SRILM
- Weights tuning : **MERT** method

Used corpora

Corpora	Designation	Size (sentences)
English-French Bilingual training		
News Commentary v6	*news-c*	116 k
Europarl v6	*euro*	1.8 M
United Nation corpus	*UN*	12 M
10^9 corpus	*giga*	23 M
English Monolingual training		
News Commentary v6	*mono-news-c*	181 k
Shuffled News Crawl corpus (from 2007 to 2011)	*news-s*	25 M
Europarl v6	*mono-euro*	1.8 M
Development		
newstest2008	*test08*	2,051
newssyscomb2009	*testcomb09*	502
newstest2009	*test09*	2,525
Test		
newstest2010	*test10*	2,489

LIG system

- Pre-processing:
  . **Lowercased** text
  . Normalization of **French euphonious 't'**

- **4-gram LM**: *mono-news-c + news-s + mono-euro*

- **Translation model**: Training on *news-c + euro + UN*

♯	System description	BLEU score: *test09*	BLEU score: *test10*
1	Training: *euro+news-c*	24.89	26.01
2	**Training:** *euro+news-c+UN*	25.44	26.43
3	2 + LM_1	24.81	27.19
4	2 + $\mathbf{LM_2}$	25.37	27.25
5	4 + **MERT** on *test09*	26.83	27.53
6	5 + **phrase-table filtering**	27.09	**27.64**
7	6 + SPE	27.53	27.74
8	6 + recaser	24.95	26.07

Table : Incremental improvement of the LIG system in terms of case-insensitive BLEU (%), except for line 8 where case-sensitive BLEU (%) are reported

500 n-best generated for each sentence + 14 decoding scores

Lowercasing of each n-bests

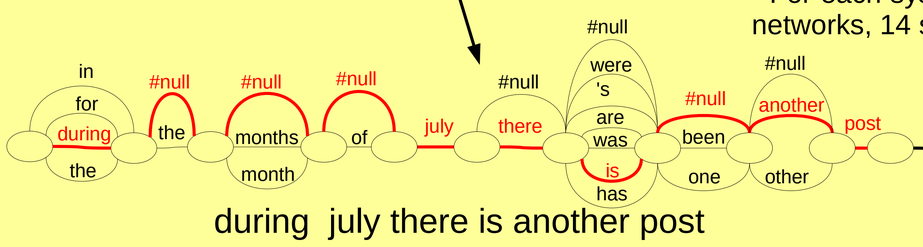
Nbest optimization : simplex-based Amoeba search on BLEU score

500 n-best generated for each sentence + 14 decoding scores

Nbest optimization : simplex-based Amoeba search on the BLEU score

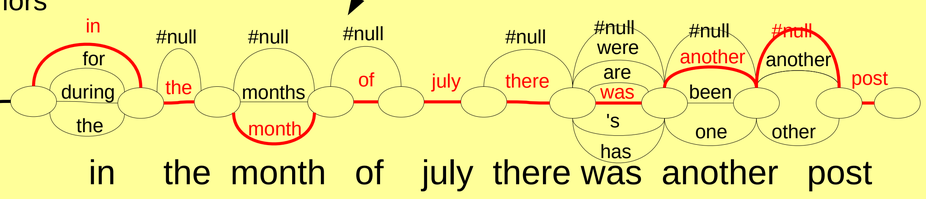
LIGA system

- For each system, n-best list are turned into confusion networks, 14 scores are used to compute posteriors



during july there is another post

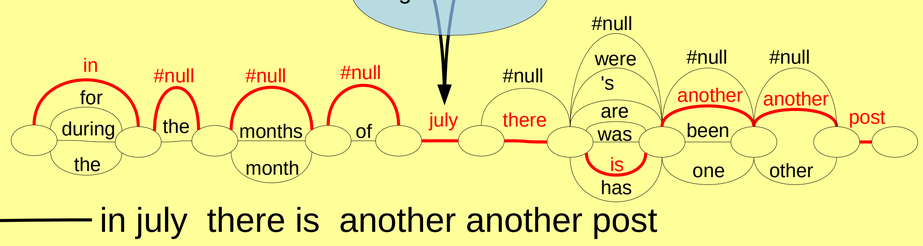
0.6



in the month of july there was another post

0.4

Merge + ROVER



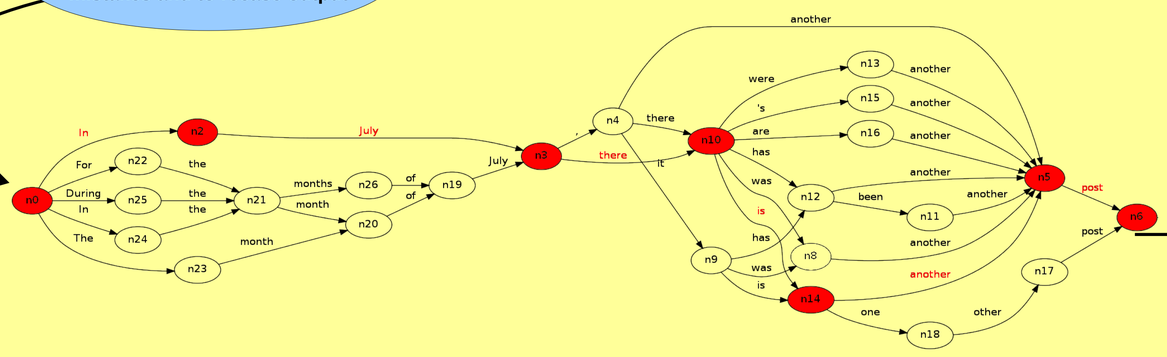
in july there is another another post

- The 500-best truecased outputs of the LIA system are first merged in a word graph

Confusion networks are merged in a single one

A ROVER is applied in order to output a one best

Oracle of the output on the graph in order to avoid some mistakes and to recase output



In July there is another post

- The new hypothesis is based on a "true" decoding pass generated by a truecased system and discarded marginal hypotheses.

Results and Conclusion

		LIG	LIA	LIG CNC	LIA CNC	LIG+LIA
case-insensitive BLEU	*test10*	27.6	29.3	28.1	29.4	29.7
	test11	28.5	29.4	28.5	29.3	29.9
case-sensitive BLEU	*test10*	26.1	28.4	27.0	28.4	28.7
	test11	26.9	28.4	27.5	28.4	28.8

Table : Performance measured before and after combining systems

Two statistical machine translation systems have been developed at different sites using MOSES and the final submission is the combination of these systems. Each individual system led to specific works.

→ **The LIGA submission presented this year was ranked among the best MT system for the French-English direction**