



HAL
open science

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

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

► **To cite this version:**

Marion Potet, Raphaël Rubino, Benjamin Lecouteux, Stéphane Huet, Hervé Blanchon, et al.. The LIGA (LIG/LIA) Machine Translation System for WMT 2011. Sixth Workshop on Statistical Machine Translation WMT 2011, 2011, Edinburgh, United Kingdom. hal-02088892

HAL Id: hal-02088892

<https://hal.science/hal-02088892v1>

Submitted on 8 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 resources (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	with
<i>news-c + euro</i> (1.77 M)	28.1	28.0
<i>news-c + 1.77 M of UN</i>	27.2	-
<i>news-c + 1.77 M of giga</i>	27.1	-
<i>news-c + 1.77 M with IR</i>	28.2	-
<i>news-c + 3 M with IR</i>	29.1	29.0
<i>news-c + 5 M with IR</i>	28.8	-
<i>news-c + 10 M with IR</i>	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

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	test10
1	Training: <i>euro+news-c</i>	24.89	26.01
2	Training: <i>euro+news-c+UN</i>	25.44	26.43
3	2 + LM_1	24.81	27.19
4	2 + LM_2	25.37	27.25
5	4 + MERT on <i>test09</i>	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

Used corpora

CORPORA	DESIGNATION	SIZE (SENTENCES)
English-French Bilingual training		
News Commentary v6	<i>news-c</i>	116 k
Europarl v6	<i>euro</i>	1.8 M
United Nation corpus	<i>UN</i>	12 M
10 ⁹ corpus	<i>giga</i>	23 M
English Monolingual training		
News Commentary v6	<i>mono-news-c</i>	181 k
Shuffled News Crawl corpus (from 2007 to 2011)	<i>news-s</i>	25 M
Europarl v6	<i>mono-euro</i>	1.8 M
Development		
newstest2008	<i>test08</i>	2,051
newssyscomb2009	<i>testcomb09</i>	502
newstest2009	<i>test09</i>	2,525
Test		
newstest2010	<i>test10</i>	2,489

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

Lowercasing of each n-bests

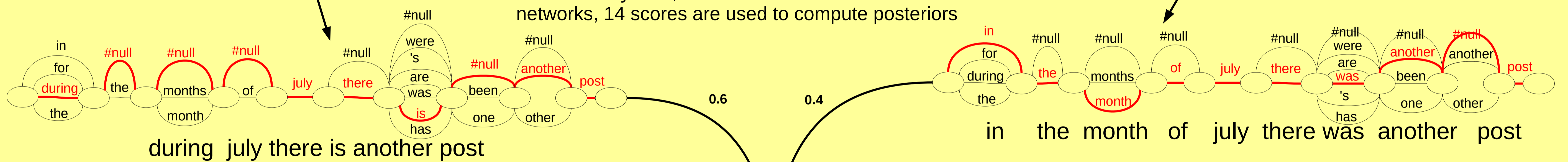
Nbest optimization : simplex-based Amoeba search on BLEU score

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



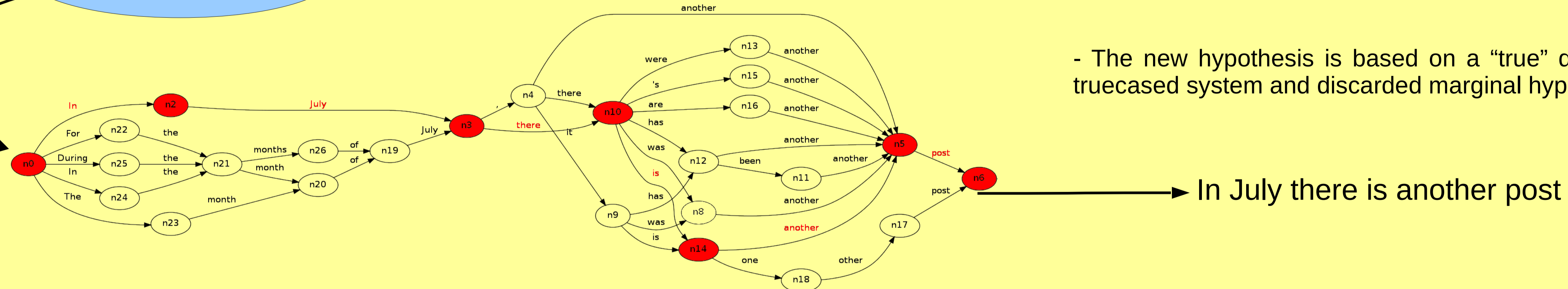
- The 500-best truecased outputs of the LIA system are first merged in a word graph

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 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 <i>test10</i>	27.6	29.3	28.1	29.4	29.7
case-insensitive BLEU <i>test11</i>	28.5	29.4	28.5	29.3	29.9
case-sensitive BLEU <i>test10</i>	26.1	28.4	27.0	28.4	28.7
case-sensitive BLEU <i>test11</i>	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