LIG System for Word Level QE task at WMT14

Ngoc Quang Luong, Laurent Besacier, Benjamin Lecouteux

LIG System for Word Level QE Task at WMT14

Ngoc-Quang Luong Laurent Besacier Benjamin Lecouteux

Laboratoire d'Informatique de Grenoble, France

INTRODUCTION

- **Task 2, WMT14:** Word-level Confidence Estimation
- **New point:** MT outputs are collected from multiple translation means (machine and human).
- **Our approach: (Figure 1)**

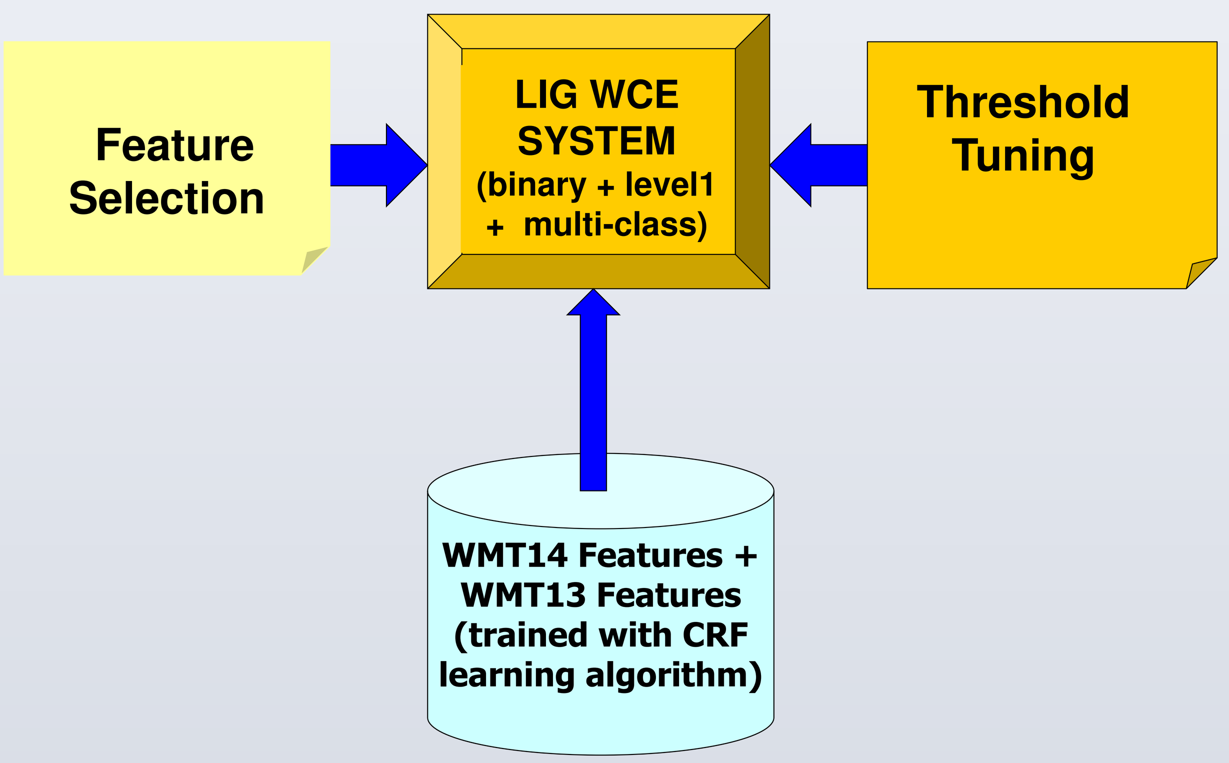

Figure 1: LIG approach for Task 2, WMT 2014

FEATURE TYPES (24 IN TOTAL)

- **The conventional features** (Table 3): work efficiently in our WMT13 submissions and are inherited in this year's systems.
- **The WMT14 features** (**bold** and *italic* in Table 3): are more specifically suggested by us for this year.

EXPERIMENTAL SETTINGS AND PRELIMINARY EXPERIMENTS

- **Data sets:** Description (Figure 2), Statistics (Table 1)

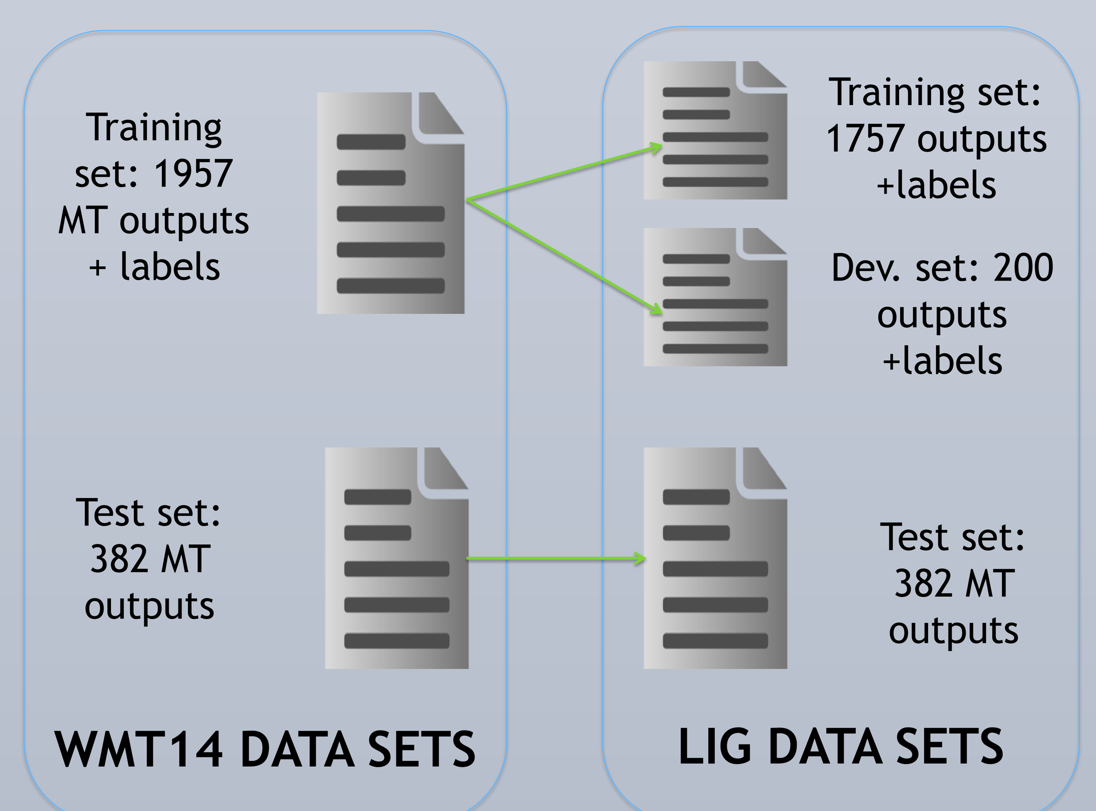

Figure 2: Data sets

Data Set	Train	Dev	Test
#segments	1757	200	382
#words	40975	6436	9613
%OK : %BAD	67 : 33	58 : 42	-
%OK: %Accuracy: %Fluency	67 : 26: 7	58 : 32: 10	-

Table 1: Statistics of training, dev and test sets.

- **WMT13 data is used to combine with WMT14 data in binary variant.**
- **Machine learning method:** Conditional Random Fields (CRF).
- **Toolkit for training and labeling:** WAPITI.
- **For binary system:** the classification corresponds to a threshold increase from 0.300 to 0.975 (step = 0.025) (Figure 3). Optimal value = 0.75.
- **Results:** BL_BIN, BL_L1, BL_MULT and BL+WMT13_BIN in Table 2.

FEATURE SELECTION (FS)

- **Objectives:** filter the most informative features, eliminate the useless ones.
- **Sequential Backward Selection**
- **Best systems :** FS_BIN, FS_L1, FS_MULT (Table 2).

Rank	Name and description
1	Target POS
2	Longest target n-gram length (the longest sequence formed by the word and the previous ones in the target LM)
3	***Occurrence in multiple systems (if the word appears in at least 50% references for the same source sentence)***
4	Target word
5	Occur in Google Translate
6	Source POS
7	Numerical (is the word numerical or not?)
8	Target Polysemy (number of senses)
9	Left source context (target word + the word before the source word aligned to it)
10	Right target context (source word + two words before the target word)
11	Constituent label (extracted from constituent tree)
12	***Longest target POS n-gram length*** *(the longest sequence formed by the word's POS and the previous ones in the target POS LM)*
13	Punctuation (is the word a punctuation?)
14	Stop word (is the word a stop-word?)
15	Number of occurrences (How many times the word appears in the sentence)
16	Left target context (source word + two words after the target word)
17	Backoff behaviour (score assigned according to how many times the target LM backs off)
18	Source Polysemy (number of senses)
19	Source Word
20	Proper name (is the word a proper name?)
21	Distance to Root (distance from this word to the root in the constituent tree
22	Longest source n-gram length (like above, but in the source LM).
23	Right Source Context (target word + the word after the source word aligned to it)
24	***Longest source POS n-gram length*** *(the longest sequence formed by the word's POS and the previous ones in the source POS LM)*

Table 3: The rank of each feature (in term of usefulness).

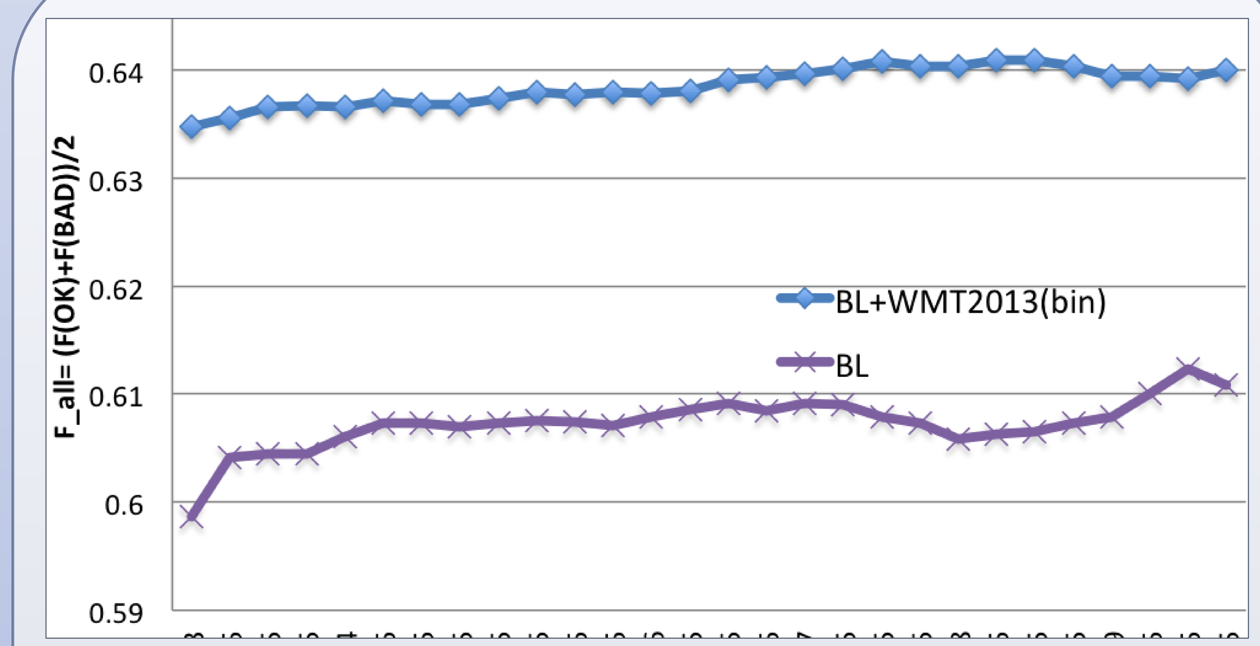

Figure 3: Decision threshold tuning on BL(bin) and BL+WMT2013(bin)

System	Label	Pr(%)	Rc(%)	F(%)
BL_BIN	OK	66.67	81.92	73.51
	BAD	60.69	41.92	49.58
BL_L1	OK	63.86	82.83	72.12
	Accuracy	22.14	14.89	17.80
	Fluency	50.40	27.98	35.98
BL_MULT	All labels	Favg(all) = 24.84		
BL+WMT13_BIN	OK	68.62	82.69	75.01
	BAD	64.38	45.73	53.47
FS_BIN	OK	68.89	83.14	75.35
	BAD	64.66	46.37	54.00
FS_L1	OK	64.03	83.47	72.47
	Accuracy	22.44	15.68	18.46
	Fluency	51.71	27.67	36.05
FS_MULT	All labels	Favg(all) = 24.88		

Table 2: Pr, Rc and F for labels of all binary , level 1 and multi-class systems, obtained on dev set.

- Performance Evolution during FS (Figure 4)
- Best subset: Top 18
- Best proposed feature: Occurrence in mult. systems

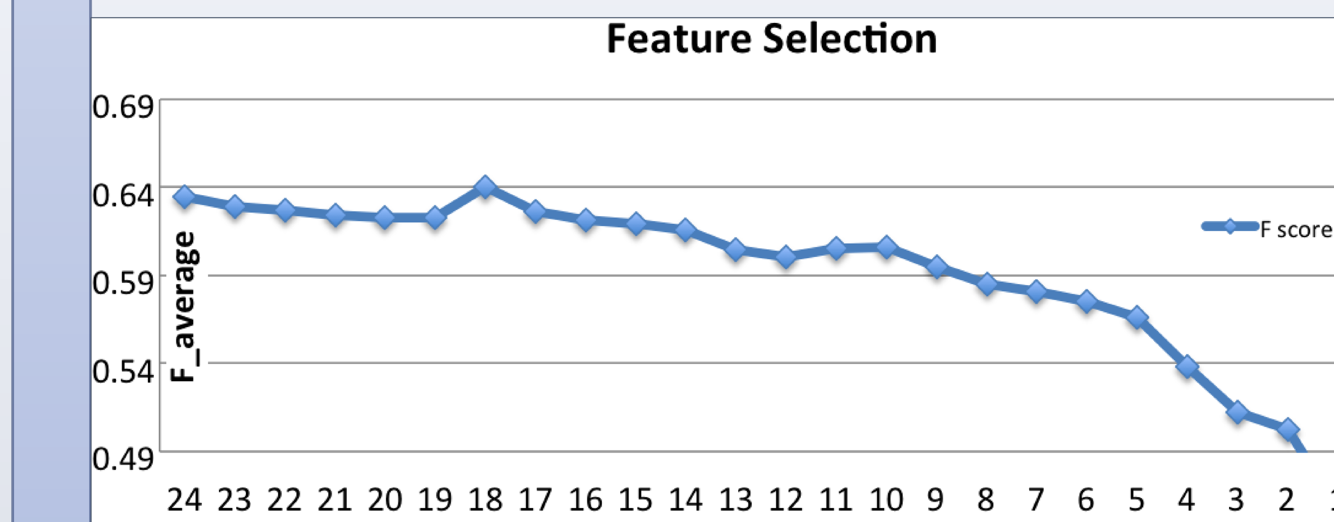

Figure 4: Evolution of system performance (Favg (all)) during Feature Selection process, obtained on dev set

SUBMISSIONS AND OFFICIAL RESULTS

- Average F(main metric): average F1 for all but the 'OK' class.
- F('OK'): F1 for 'OK' class.

System	Average F(%)	F('OK') (%)
FS_BIN (primary)	44.4735	74.0961
FS_L1	31.7814	73.9856
FS_MULT	20.4953	76.6645
BL+WMT13(BIN)	44.1074	74.6503
BL_L1	31.7894	74.0045
BL_MULT	20.4953	76.6645

Table 4: Official results of the submitted systems.

CONCLUSIONS AND PERSPECTIVES

- Integration of several novel features.
- Feature Selection's help to enlighten the valuable features.
- More data (WMT13) helps to boost performance
- Future work: research linguistic features, reinforce the segment- level CE, propose the methodology for Sentence CE relied on the word- and segment- level.