An Open Source Toolkit for Word-level Confidence Estimation in Machine Translation

Christophe Servan, Ngoc Tien Le, Ngoc Quang Luong, Benjamin Lecouteux and Laurent Besacier

4 Dec. 2015

Outline

Confidence Estimation (CE)

How much I trust the MT sytem outputs

Confidence Measures for Machine Translation

- Several levels:
  - Word
  - Sentence
  - Document
- Classification task:
  - Binary label (*e.g.*: Good/Bad)
  - Multiple labels (*e.g.*: kind of error)

  $\Rightarrow$ can produce a score associated with the label.

$\Rightarrow$ Word-level Confidence Estimation with binary labels (and scores)

Framework

Computer Aided Translation (CAT) Tool

- help professionnal translators
- highlight which part to focus on
- speed up the post-editing process

$\Rightarrow$ already part of several CAT tools CasmaCAT/MateCAT, libellex (Lingua & Machina)...

Toolkit overview

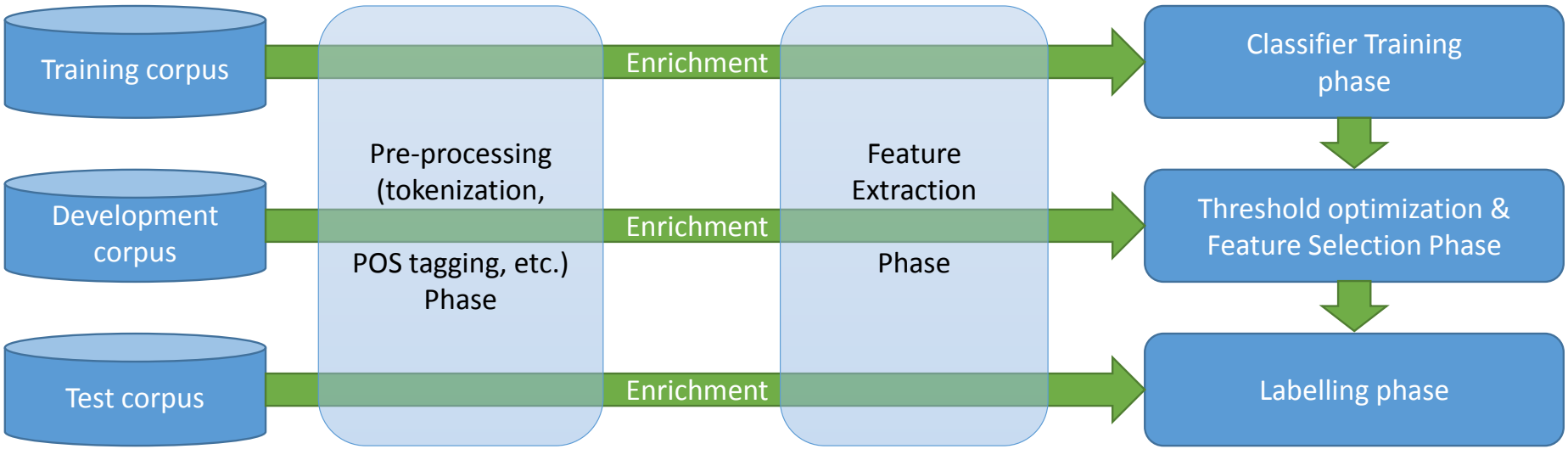

Figure: Pipeline of our Word-level Confidence Estimation tool

- Reproductible expriments (WMT14 QE shared task)
- Easy to add new features (written in python)
- External toolkits used: GIZA++, TreeTagger, SRILM, BabelNet, Berkeley parser, Wapiti...

Features Extracted

1 *Proper Name*
2 **Unknown Stem**
3 # of Word Occurrences
4 **# of Stem Occurrences**
5 Source POS
6 *Source Word*
7 **Source Stem**
8 Target POS
9 Target Word
10 **Target Stem**
11 *Word context Alignements*
12 *POS context Alignements*
13 **Stem context Alignements**
14 Longest Target *N*-gram Length
15 Longest Source *N*-gram Length
16 WPP Exact*
17 WPP Any*
18 WPP Min*
19 WPP Max*
20 Nodes*
21 Numerical
22 Punctuation
23 Stop Word
24 Target Backoff Behaviour
25 *Constituent Label*
26 *Distance To Root*
27 *Polysemy Count – Target*
28 **Occur in Bing Translator**
29 Occur in Google Translate

*: internal features

Tasks for Word-level Confidence Estimation

two sets of expriments:

- en–es:
  - WMT14 QE shared task
  - State-of-the-Art evaluation task
  - MT outputs only (*i.e.*: 1-best only)
- fr–en:
  - in-house data
  - SMT outputs with *N*-best list

Task protocol:

- Word-level Confidence Estimation (one label per word)
- Binary labels (*OK* or *BAD*)
- Precision, Recall, F-Measure for each label and Mean F-Measure.

Amount of data (# lines):

Corpus	en–es	fr–en
Train	650	10K
Test	208	881
N-Best	1	1000

State-of-the-Art: WMT14 Quality Estimation task

The *en–es* Word-level Quality Estimation task[1]

	Systems	**M-F**	**F**(*bad*)
	FBK-UPV-UEDIN-1	62.00	48.73
	LIMSI	60.55	47.32
→	*Our toolkit*	*60.76*	*47.17*
	LIG-1	63.55	44.47
	LIG-2	63.77	44.11
	FBK-UPV-UEDIN-2	62.17	42.63

[1] Without internal features

Decision threshold optimization (oracle)

Task	Threshold	Label	P	R	F	M-F
en–es (WMT14)	*Default*	*Good*	71.24	**77.73**	**74.35**	60.76
		Bad	**51.82**	43.28	47.17	
	Optimized	*Good*	**71.42**	76.82	74.03	**60.87**
		Bad	51.49	**44.45**	**47.71**	
fr–en	*Default*	*Good*	84.45	90.22	**87.24**	64.96
		Bad	**50.10**	37.16	42.67	
	Optimized	*Good*	**85.60**	85.65	85.62	**65.59**
		Bad	45.61	**45.50**	**45.56**	

Set of features (no threshold optimization)

The *fr–en* task only:

Features	Labels	P	R	F	M-F
Base.	*Good*	81.97	**92.22**	86.80	58.64
	Bad	44.17	23.28	30.48	
+ *mod.*	*Good*	83.21	90.99	86.92	62.00
	Bad	47.24	30.53	37.09	
+ *new*	*Good*	83.55	90.11	86.70	62.65
	Bad	46.75	32.86	38.60	
+ *MT*	*Good*	**84.45**	90.22	**87.24**	**64.96**
	Bad	**50.10**	**37.16**	**42.67**	

Sequential Forward Selection algorithm (*fr–en* task only)

Rank	Feature	Rank	Feature
1	**Stem context Alignements**	16	Stop Words
2	WPP Exact	17	Nodes
3	*Word context Alignements*	18	**# of Stem Occurrences**
4	WPP Max	19	Numeric
5	WPP Any	20	**Unknown Stem**
6	WPP Min	21	Target Word
7	*POS context Alignements*	22	Source POS
8	Occur in Google Translate	23	*Polysemy Count – Target*
9	Longest Target *N*-gram Length	24	*Source Word*
10	**Occur in Bing Translate**	25	*Constituent Label*
11	**Source Stem**	26	Punctuation
12	Target Backoff Behaviour	27	**Target Stem**
13	Longest Source *N*-gram Length	28	*Proper Name*
14	# of Word Occurrences	29	Target POS
15	*Distance To Root*		

Best Mean F-Measure score: 65.14 points.

GitHub exploration:

`https://github.com/besacier/WCE-LIG`

Conclusion and perspectives

Contributions:

- Open-Source, Flexible and State-of-the-Art WCE Toolkit
- Decision Threshold and Feature Selection (SFS)
- Available on GitHub:
  `https://github.com/besacier/WCE-LIG`

Perspectives:

- More flexible!
- Feature Selection with
  Sequential Floating Forward Selection algorithm
- Add new features based on Word Embeddings
  (monolingual and bilingual)
- Propose a version of the WCE toolkit for Speech Translation

Thanks for your attention!

Remarks & questions are welcome!

contact: christophe.servan@imag.fr & laurent.besacier@imag.fr