MACHINE TRANSLATION QUALITY ESTIMATION

A Linguist's Approach

ebay

WHAT IS MT QUALITY ESTIMATION?

Automatically providing a quality indicator for machine translation output without depending on human reference translations.

Our objective:
Estimate quality and post-editing effort for eBay listing titles and descriptions

ONE big CHALLENGE

$$\min_W \sum_{t=1}^{T} ||(W^{(t)}X^{(t)} - Y^{(t)})||2^2 + \lambda_s||S||_1 + \lambda_b||B||_{1,\infty} \text{ subject to: } W = S + B$$

or

"State-of-the-art QE explores different supervised linear or non-linear learning methods for regression or classification such as Support Vector Machines (SVM), different types of Decision Trees, Neural Networks, Elastic-Net, Gaussian Processes, Naive Bayes, among others"

(Machine Translation Quality Estimation Across Domains, de Souza et al, year))

A LINGUIST'S APPROACH

Using linguistic features from 3 dimensions:

COMPLEXITY **ADEQUACY** **FLUENCY**

FEATURES

Complexity:

- Length
- Polysemy

Adequacy:

- QA
  - Terminology
  - Patterns
  - Blacklist
  - Numbers
- Automated Post-Editing
- (POS)
- (NER)

NWT EUC Lot of 4 JCrew Old Navy Skinny Tank Tops Pink Green Grey Women's Med

$7.50 3 bids

Fluency:

- Misspellings
- Grammar errors

New Toshoba C50-A-053 Celeron 1.9Ghz 6GB DDR3 500GB 15.6" DVD HDMI USB 3.0 Si

IMPLEMENTATION

Checkmate+LanguageTool

Reusable Profile

Detailed Report

Score

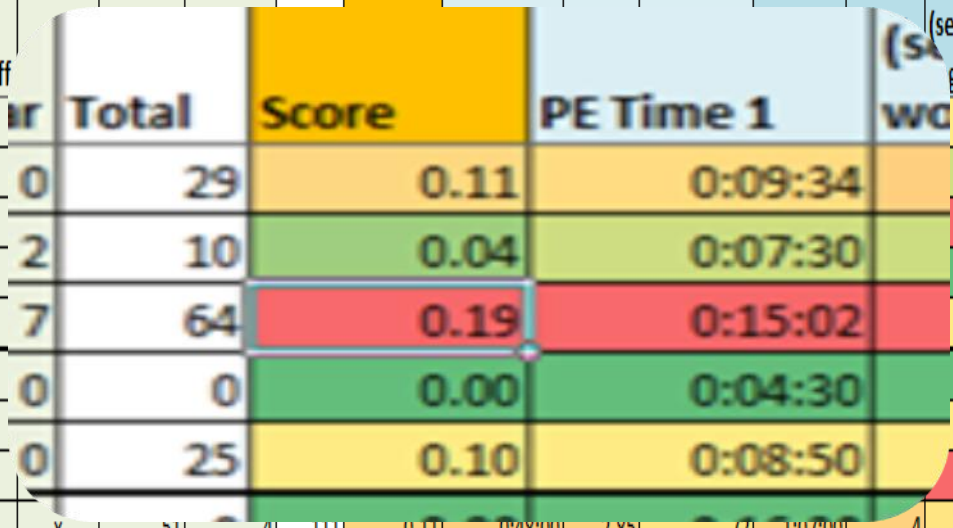
TESTING

- One Language (es-LA)
- Short samples (~300 words)
- Bigger samples (~1000 words)
- Post-Edited files (~50,000 words)
- pt-BR, ru-RU, zh-CN

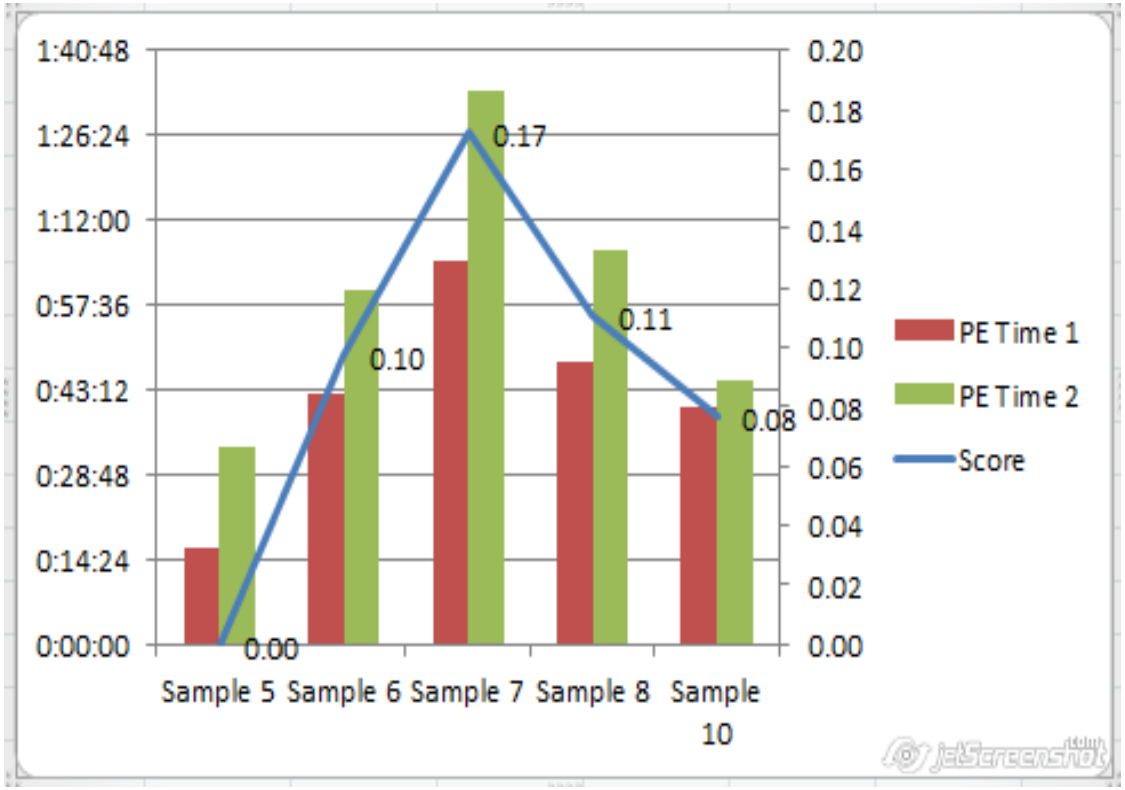
RESULTS

MEASURING RESULTS

	Statistic Data			Complexity		Adequacy				Fluency			Score	PE1			PE2		
Sample	Size (seg)	Size (w)	Aver. Seg (W/S)	Length	Polysemic	QA	Postprocessing Script changes	PoS Diff							PE Time	Ave PE time		PE Time	Ave PE time (secs x gment)
Sample 1 -	10	254	25	0	10	1	0	x											84
Sample 2 -	10	256	25	0	4	0	0	x											78
Sample 3 -	10	330	33	10	10	6	3	x											94.4
Sample 4	10	273	27	0	0	0	0	x											63.1
Sample 9	10	250	25	0	9	1	0	x											84
Sample 5	37	997	27	0	0	0	0	x											54.2
Sample 6	37	984	26	0	14	5	3	x											97.2
Sample 7	34	1033	30	10	31	10	1	x											165
Sample 8	40	1004	25	0	16	6	1	x	x	51	4	111	0.11	0:48:00	2.85	72	1:07:00	4	100.5
Sample 10	38	1015	26	0	20	6	3	x	x	14	4	78	0.08	0:40:30	2.39	63.9	0:45:00	2.66	



SAMPLES - SCORE AND TIME ALIGN



Sample	Score	PE Time 1	PE Time 2
Sample 5	0.00	0:16:30	0:33:27
Sample 6	0.10	0:42:45	1:00:00
Sample 7	0.17	1:05:00	1:34:00
Sample 8	0.11	0:48:00	1:07:00
Sample 10	0.08	0:40:30	0:45:00

FILES - SCORE AND ED ALIGN

XLIFF_66497_file6	Size (w)	Total	Score	Edit Distance
MT output	53777	4559	0.08478	
delivery 1	53777	3503	0.06514	64.06
delivery 2	53777	3484	0.06479	70.76

Average ED (es-LA, descriptions) = 72

MT QE OVER TIME

Sample	Size (w)	Total	Score	Date
mt_desc_items_latam_2014_02_19.400k.final_train_WLA1	47728	5558	0.11645	3/2/2014
mt_desc_items_latam_2014_02_19.400k.final_train_WLA2	46746	6135	0.13124	3/2/2014
mt_desc_items_latam_2014_02_19.400k.final_train_WLA3	47743	5814	0.12178	3/2/2014
mt_desc_items_latam_2014_02_19.400k.final_train_WLA4	46558	6054	0.13003	3/2/2014
66497 MT_descriptions_training_PE_ENESLA_File1	56779	5237	0.09223	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File2	56470	5475	0.09695	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File3	56714	5726	0.10096	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File4	56819	5546	0.09761	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File5	56286	5253	0.09333	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File6	53777	4559	0.08478	4/25/2014
68865_en-es-CO-descriptions-mar15-b-translated.xliff_0	50433	3640	0.07217	5/3/2015
1293_MT_eBay_descriptions_en_es_latam_30k_PE_P2-B_0	32209	3120	0.09687	7/2/2015
1294_MT_eBay_descriptions_en_es_latam_70k_PE_P1-B_1	21500	2266	0.10540	7/2/2015

SAMPLES - OTHER LANGUAGES

	Words	Issues	Score	PE Time
BPT				
PE Sample 1	500	21	0.0420	33 min
PE Sample 2	500	11	0.0220	14 min
PE Sample 3	500	7	0.0140	5 min
RU				
Sample 1	500	91	0.1820	58 min
Sample 2	500	75	0.1500	19.5 min
Sample 3	500	62	0.1240	9.3 min
ZHCN				
Sample 1	500	9	0.0180	39 min
Sample 2	500	8	0.0160	21 min
Sample 3	500	6	0.0120	15 min

CHALLENGES

- False positives
- Matching score and post-editing effort
- Same weight for all features

WHAT'S NEXT

- Tracking scores over time
- Adding scores to our post-editing tool
- Adding new languages
- Researching new features

HOW CAN *YOU* USE THIS?

- Tailor the model to your needs
- Estimate quality at the file/segment level
- Target post-editing, discard bad content
- Estimate post-editing effort/time
- Compare MT systems
- Monitor MT system progress

Q&A

THANK YOU!

jrowda@ebay.com