Bootstrapping Quality Estimation in a live production environment

EAMT 2017

CROSSLANG
TRANSLATION AUTOMATION

Introduction

Quality Estimation

"The process of scoring Machine Translation (MT) output without access to a reference translation"

- QE aims:
  - Hide "bad MT Output" during the Post-Editing phase
  - Take away frustration at the side of translators
  - Increase acceptance of MT + Post-Editing
- This talk:
  - Sentence-based QE, scoring (not ranking), supervised learning
  - Summary of a one-year project

CROSSLANG
TRANSLATION AUTOMATION

Project context

Different aims in academia and industry

- In academia:
  - development/testing of algorithms and features to better learn estimates
- In industry:
  - come to a workable real-time solution
  - define best practices
  - find workarounds for limiting factors (this talk: “bootstrapping” by lack of Post-Edits to learn from)
  - productize knowledge (MT + QE score)

CROSSLANG
TRANSLATION AUTOMATION

Outline

- Our implementation
  - How QE should have been done,
    according to the research literature *(estimating Post-Edit distance)*
  - Project constraints
  - How it was done,
    considering the constraints (estimating Post-Edit Effort judgment scores)
  - Results
- Validation
  - Compare ***PE effort judgment score*** prediction to ***PE distance*** prediction
- Further experiments

CROSSLANG
TRANSLATION AUTOMATION

Implementation

CROSSLANG
TRANSLATION AUTOMATION

WMT 2013 protocol

- Predicting PE distance
- HTER distance [0 … 1] as labels
  - HTER: perform the minimum number of post-editing operations to obtain acceptable output
  - “Minimum PE” versus reference translation: easier to predict
  - Eliminate subjectivity of effort judgment scores
  - Eliminate variance in effort judgment scores
  - Disadvantage: “Minimum PE” vs. production quality PE

CROSSLANG
TRANSLATION AUTOMATION

Project context/constraints

- 9 Phrase-Based SMT systems for 3 domains (IT-related), sizes: see table
- Not released for production yet
- No Post-Edits available (except for DOM1 EN-DE)
- HTER post-edits considered to be wasteful

DOMAIN	DOM1	DOM2	DOM3
DE-EN	2,613,489	22,375,900	-
EN-DE	2,971,501	13,838,326	1,154,653
EN-ZH	-	2,557,042	439,980
EN-ES	-	3,456,275	366,423
EN-PT	-	2,942,499	298,687
EN-FR	-	4,944,361	343,352
EN-RU	-	2,108,723	455,203
EN-IT	-	3,198,050	-
EN-JP	878,036	4,915,823	533,053

CROSSLANG
TRANSLATION AUTOMATION

Simplified WMT 2012 protocol
PE Effort judgments

WMT 2012

- Human PE effort judgments
  - Non-professional translators
  - Intra-annotator agreement (control group of repeated annotations)
  - Data discarded
- Scoring task
  - Present source + MT output + post-edit
- Score weighting

Our approach

- Human PE effort judgments
  - Professional translators
  - Only inter-annotator agreement
  - All data preserved
- Scoring task
  - Present source + MT output
- Score weighting

CROSSLANG
TRANSLATION AUTOMATION

Simplified WMT 2012 protocol Scores

WMT 2012 — Our approach

1. The MT output is ***incomprehensible***, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.
2. About ***50-70% of the MT output needs to be edited***. It requires a significant editing effort in order to reach publishable level.
3. About ***25-50% of the MT output needs to be edited***. It contains different errors and mistranslations that need to be corrected.
4. About ***10-25% of the MT output needs to be edited***. It is generally clear and intelligible.
5. The ***MT output is perfectly clear and intelligible***. It is not necessarily a perfect translation but requires little or no editing.

Resulting data set

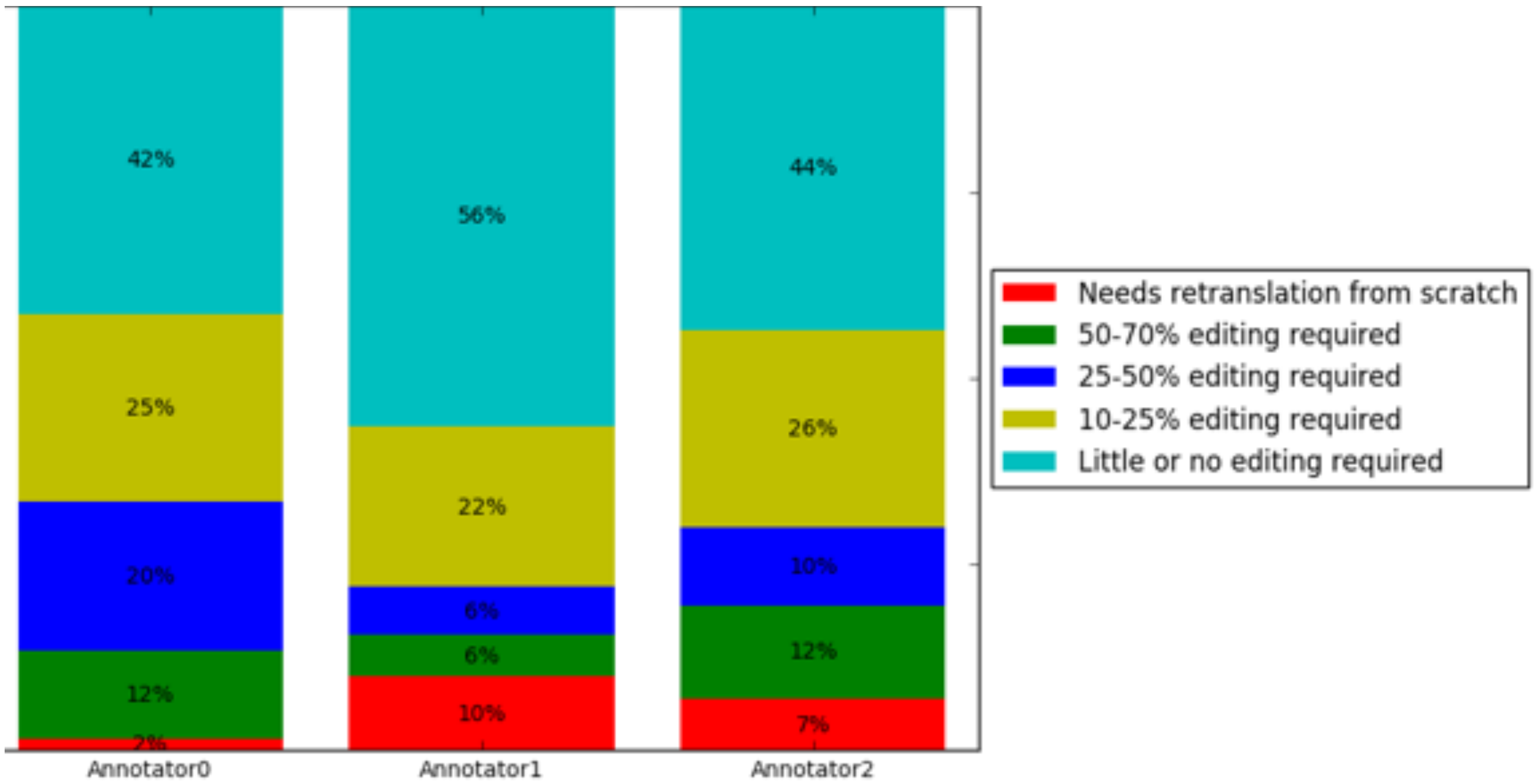
- 800 sentences
- 3 professional annotators
- DOM1 underrepresented, but it is the only domain for which we have Pes (EN-DE)

DOMAIN	DOM1	DOM2	DOM3	TOTAL
DE-EN	800	800	-	1,600
EN-DE	800	800	800	2,400
EN-ZH	-	800	800	1,600
EN-ES	-	800	800	1,600
EN-PT	-	800	800	1,600
EN-FR	-	800	800	1,600
EN-RU	-	800	800	1,600
EN-IT	-	800	-	800
EN-JP	800	800	800	2,400

CROSSLANG
TRANSLATION AUTOMATION

Resulting data set

- MT output already reasonable good
- Inter-annotator agreement *fair*, at 0.44 Fleiss' coefficient



CROSSLANG
TRANSLATION AUTOMATION

Results

QE systems trained

- for each data set, ***language + domain-specific*** models were trained (listed in the white columns)
- ***language-specific*** models were trained by combining all data available for each language pair (listed in the white LANG row).
- language agnostic ***domain-specific*** models were trained by aggregating all data for each domain separately (ALL column in grey).
- finally, a language-agnostic **BULK** model (BULK row in grey), with all available data was trained.

CROSSLANG
TRANSLATION AUTOMATION

Focus on deployment configurations

Domain MAE/MRSE	De-En		En-De		En-Zh		En-Es		En-Pt		En-Fr		En-It		All	
Dom1	0.65	0.88	0.68	0.88	-		-		-		-		-		0.73	0.97
Dom2	0.54	0.86	0.94	1.16	0.79	1.06	0.63	0.98	0.77	0.99	0.54	0.76	0.62	0.87	0.76	1.03
Dom3	-		0.80	1.05	0.68	0.95	0.54	0.85	0.86	1.10	0.63	0.95	-		0.79	1.03
Lang	0.63	0.90	0.80	1.03	0.70	0.97	0.52	0.83	0.76	1.02	0.55	0.80	0.62	0.87	0.77	1.04
Bulk	0.77								1.04							

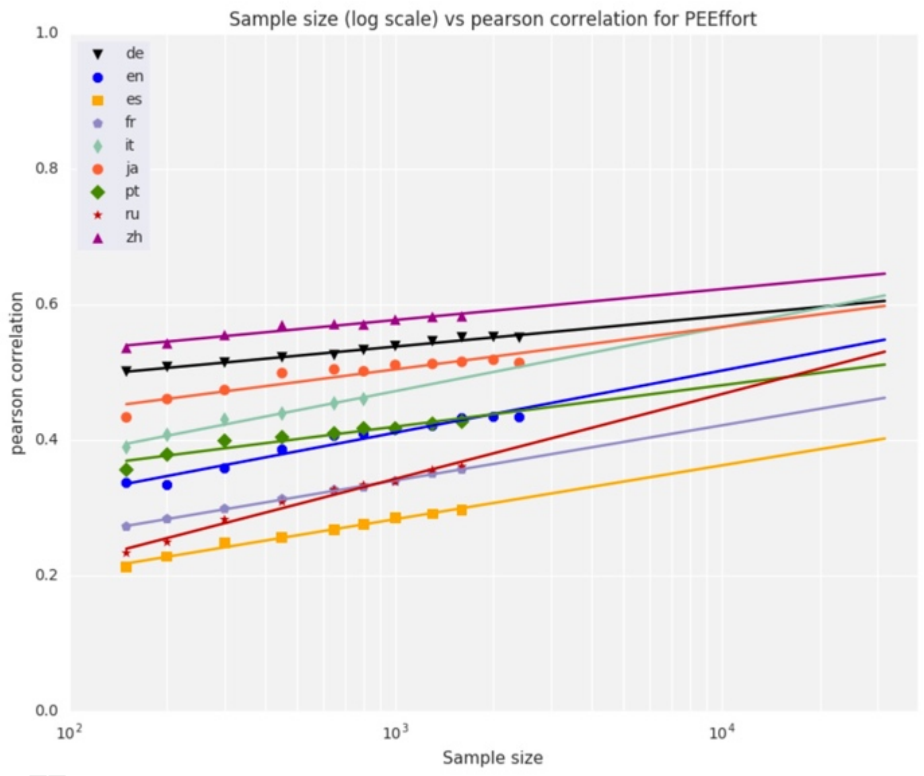
CROSSLANG
TRANSLATION AUTOMATION

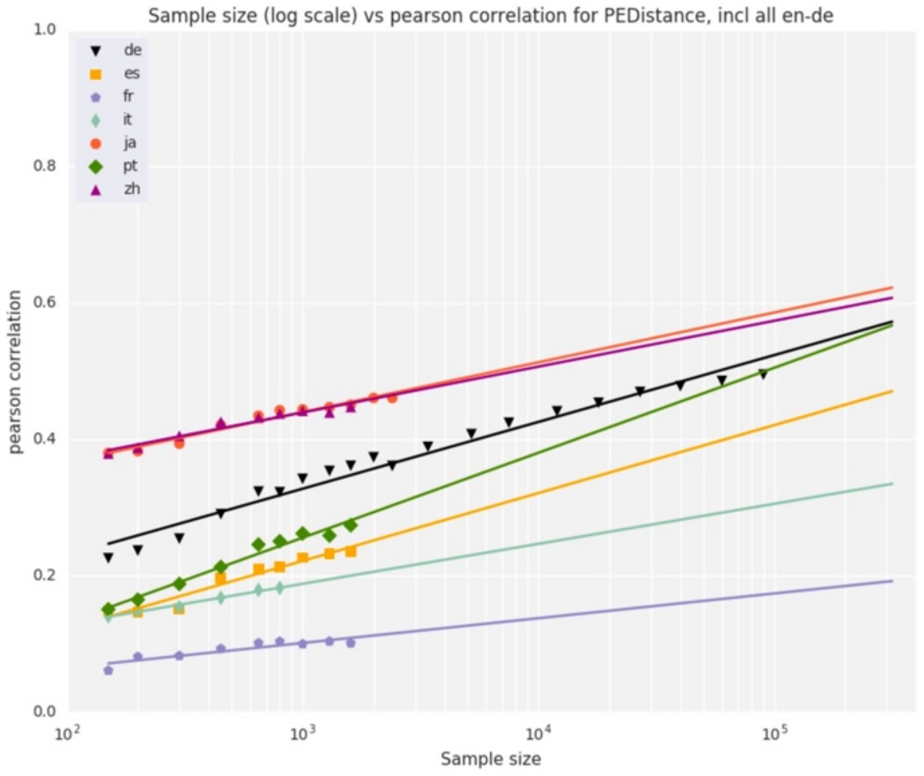
Validation
of our approach

Motivation

- Assume: 800 PE judgment (x3) as expensive as actual PE
- Question: Is our system better than a system based on 2,400 PE distance labels?
- Caveats:
  - PE effort [0 .. 5] vs. PE distance [0 … 1], Pearson correlation as go-between
  - PE distance more difficult to predict on reference translations (easier on “Minimum PEs”)

PE effort judgments vs. PE distance





CROSSLANG
TRANSLATION AUTOMATION

Further experiments

Technical OOVs

- example: ecl_kd042_de_crm_basis (Fishel & Sennrich 2014)
- technical OOVs are normalized. If this behavior is not compensated for by the QE system, sentences with technical OOVs will unrightfully receive a penalty at lookup time
- technical OOVs, require a simple copy operation (if not resolved by the MT system), which makes the task of sentences containing OOVs easier, instead of more difficult
- custom classifier for Technical OOVs

CROSSLANG
TRANSLATION AUTOMATION

Web-Scale LM & Syntactic Features

- Yandex paper (Kozlova et al., 2016), using SyntaxNet (Andor, et al., 2016)
- Tree-based features
  (tree width, maximum tree depth, average tree depth, …)
- Features derived from Part-Of-Speech (POS) tags and dependency roles (number of verb, number of verbs with dependent subjects, number of nouns, number of subjects, number of conjunctions, number of relative clauses, …)
- Experiments were run on the EN-DE PE distance data set

Results PE distance labels, with reference translation

Sample Size	Features Set	#	Mae	Pearson Correlation
700	Baseline	19	0.27+/-0.01	0.26+/-0.02
	+ Syntax	43	0.26+/-0.01	0.32+/-0.01
	+ Syntax + WebLM	45	0.27+/-0.01	0.32+/-0.01
7,000	Baseline	19	0.24+/-0.01	0.43+/-0.01
	+ Syntax	43	0.24+/-0.01	0.46+/-0.01
	+ Syntax + WebLM	45	0.24+/-0.01	0.46+/-0.01
70,000	Baseline	19	0.23+/-0.01	0.50+/-0.01
	+ Syntax	43	0.22+/-0.01	0.55+/-0.01
	+ Syntax + WebLM	45	0.22+/-0.01	0.56+/-0.01

CROSSLANG
TRANSLATION AUTOMATION

Conclusions

CROSSLANG
TRANSLATION AUTOMATION

PE effort judgments still useful?

- “Cheap” alternative to “wasteful” Post-Edits that do not meet production quality guidelines
- Can create a baseline when searching optimum data split between MT training/QE training (in large (+10M sentence pairs) MT environments)
- Can create a baseline to get an idea of the required data set size for PE distance based QE
- Comparison PE effort judgments and PE distance should be improved

CROSSLANG
TRANSLATION AUTOMATION