Bootstrapping
Quality Estimation in a
live production environment

EAMT 2017
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
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)
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
Implementation
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
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

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dom1</th>
<th>Dom2</th>
<th>Dom3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE-EN</td>
<td>2,613,489</td>
<td>22,375,900</td>
<td>-</td>
</tr>
<tr>
<td>EN-DE</td>
<td>2,971,501</td>
<td>13,838,326</td>
<td>1,154,653</td>
</tr>
<tr>
<td>EN-ZH</td>
<td>-</td>
<td>2,557,042</td>
<td>439,980</td>
</tr>
<tr>
<td>EN-ES</td>
<td>-</td>
<td>3,456,275</td>
<td>366,423</td>
</tr>
<tr>
<td>EN-PT</td>
<td>-</td>
<td>2,942,499</td>
<td>298,687</td>
</tr>
<tr>
<td>EN-FR</td>
<td>-</td>
<td>4,944,361</td>
<td>343,352</td>
</tr>
<tr>
<td>EN-RU</td>
<td>-</td>
<td>2,108,723</td>
<td>455,203</td>
</tr>
<tr>
<td>EN-IT</td>
<td>-</td>
<td>3,198,050</td>
<td>-</td>
</tr>
<tr>
<td>EN-JP</td>
<td>878,036</td>
<td>4,915,823</td>
<td>533,053</td>
</tr>
</tbody>
</table>
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
Simplified WMT 2012 protocol Scores

WMT 2012

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.

Our approach
Resulting data set

- 800 sentences
- 3 professional annotators
- Dom1 underrepresented, but it is the only domain for which we have Pes (EN-DE)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dom1</th>
<th>Dom2</th>
<th>Dom3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE-EN</td>
<td>800</td>
<td>800</td>
<td>-</td>
<td>1,600</td>
</tr>
<tr>
<td>EN-DE</td>
<td>800</td>
<td>800</td>
<td>800</td>
<td>2,400</td>
</tr>
<tr>
<td>EN-ZH</td>
<td>-</td>
<td>800</td>
<td>800</td>
<td>1,600</td>
</tr>
<tr>
<td>EN-ES</td>
<td>-</td>
<td>800</td>
<td>800</td>
<td>1,600</td>
</tr>
<tr>
<td>EN-PT</td>
<td>-</td>
<td>800</td>
<td>800</td>
<td>1,600</td>
</tr>
<tr>
<td>EN-FR</td>
<td>-</td>
<td>800</td>
<td>800</td>
<td>1,600</td>
</tr>
<tr>
<td>EN-RU</td>
<td>-</td>
<td>800</td>
<td>800</td>
<td>1,600</td>
</tr>
<tr>
<td>EN-IT</td>
<td>-</td>
<td>800</td>
<td>-</td>
<td>800</td>
</tr>
<tr>
<td>EN-JP</td>
<td>800</td>
<td>800</td>
<td>800</td>
<td>2,400</td>
</tr>
</tbody>
</table>
Resulting data set

- MT output already reasonable good
- Inter-annotator agreement *fair*, at 0.44 Fleiss’ coefficient
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.
Focus on deployment configurations

<table>
<thead>
<tr>
<th>DOMAIN MAE/MRSE</th>
<th>DE-En</th>
<th>EN-DE</th>
<th>EN-ZH</th>
<th>EN-ES</th>
<th>EN-PT</th>
<th>EN-FR</th>
<th>EN-IT</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dom1</td>
<td>0.65</td>
<td>0.88</td>
<td>0.68</td>
<td>0.88</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dom2</td>
<td>0.54</td>
<td>0.86</td>
<td>0.94</td>
<td>1.16</td>
<td>0.79</td>
<td>1.06</td>
<td>0.63</td>
<td>0.98</td>
</tr>
<tr>
<td>Dom3</td>
<td>-</td>
<td>0.80</td>
<td>1.05</td>
<td>0.68</td>
<td>0.95</td>
<td>0.54</td>
<td>0.85</td>
<td>0.86</td>
</tr>
<tr>
<td>Lang</td>
<td>0.63</td>
<td>0.90</td>
<td>0.80</td>
<td>1.03</td>
<td>0.70</td>
<td>0.97</td>
<td>0.52</td>
<td>0.83</td>
</tr>
<tr>
<td>Bulk</td>
<td>0.62</td>
<td>0.86</td>
<td>0.62</td>
<td>0.87</td>
<td>0.62</td>
<td>0.87</td>
<td>0.77</td>
<td>1.04</td>
</tr>
</tbody>
</table>

**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”)

CROSSLANG
TRANSLATION AUTOMATION
PE effort judgments vs. PE distance
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
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

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Features Set</th>
<th>#</th>
<th>Mae</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>700</td>
<td>Baseline</td>
<td>19</td>
<td>0.27+/−0.01</td>
<td>0.26+/−0.02</td>
</tr>
<tr>
<td>7,000</td>
<td>+ Syntax</td>
<td>43</td>
<td>0.26+/−0.01</td>
<td>0.32+/−0.01</td>
</tr>
<tr>
<td></td>
<td>+ Syntax + WebLM</td>
<td>45</td>
<td>0.27+/−0.01</td>
<td>0.32+/−0.01</td>
</tr>
<tr>
<td>70,000</td>
<td>Baseline</td>
<td>19</td>
<td>0.24+/−0.01</td>
<td>0.43+/−0.01</td>
</tr>
<tr>
<td></td>
<td>+ Syntax</td>
<td>43</td>
<td>0.24+/−0.01</td>
<td>0.46+/−0.01</td>
</tr>
<tr>
<td></td>
<td>+ Syntax + WebLM</td>
<td>45</td>
<td>0.24+/−0.01</td>
<td>0.46+/−0.01</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>19</td>
<td>0.23+/−0.01</td>
<td>0.50+/−0.01</td>
</tr>
<tr>
<td></td>
<td>+ Syntax</td>
<td>43</td>
<td>0.22+/−0.01</td>
<td>0.55+/−0.01</td>
</tr>
<tr>
<td></td>
<td>+ Syntax + WebLM</td>
<td>45</td>
<td>0.22+/−0.01</td>
<td>0.56+/−0.01</td>
</tr>
</tbody>
</table>
Conclusions
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