Estimating machine translation quality

State-of-the-art systems and open issues

Lucia Specia

University of Sheffield
l.specia@sheffield.ac.uk

6 September 2012
1. Quality Estimation

2. Shared Task

3. Open issues

4. Conclusions
Quality estimation (QE): metrics that provide an estimate on the quality of unseen translated texts
Quality estimation (QE): metrics that provide an estimate on the quality of unseen translated texts

Quality = Can we publish it as is?
Quality estimation (QE): metrics that provide an estimate on the quality of unseen translated texts

- Quality = Can we publish it as is?
- Quality = Can a reader get the gist?
**Quality estimation** (QE): metrics that provide an *estimate* on the *quality* of unseen translated texts

- **Quality =** Can we publish it as is?
- **Quality =** Can a reader get the gist?
- **Quality =** Is it worth post-editing it?
Overview

**Quality estimation** (QE): metrics that provide an estimate on the quality of unseen translated texts

- Quality = **Can we publish it as is?**
- Quality = **Can a reader get the gist?**
- Quality = **Is it worth post-editing it?**
- Quality = **How much effort to fix it?**
Framework

QE system

Examples: source & translations, quality scores

Quality indicators
Framework

- Source text
- MT system
- Translation
- QE system
- Quality indicators
- Quality score
- Examples: source & translations, quality scores

Estimating machine translation quality
No access to reference translations: supervised machine learning techniques to **predict** quality scores
Also called **confidence estimation**, started in 2002/3
- Inspired by confidence scores in ASR: word posterior probabilities
Background

- Also called confidence estimation, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A “hard to beat” baseline: MT is always bad
  - Poor results, no use in applications
Also called **confidence estimation**, started in 2002/3
- Inspired by confidence scores in ASR: word posterior probabilities

JHU Workshop in 2003
- Estimate BLEU/NIST/WER: difficult to interpret
- A “hard to beat” baseline: **MT is always bad**
- Poor results, no use in applications

New surge in interest from 2008/9
- Better MT systems
Also called confidence estimation, started in 2002/3
- Inspired by confidence scores in ASR: word posterior probabilities

JHU Workshop in 2003
- Estimate BLEU/NIST/WER: difficult to interpret
- A “hard to beat” baseline: MT is always bad
- Poor results, no use in applications

New surge in interest from 2008/9
- Better MT systems ✓
Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A “hard to beat” baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems ✓
  - MT used in translation industry
Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A “hard to beat” baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems ✓
  - MT used in translation industry ✓
Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A “hard to beat” baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems ✓
  - MT used in translation industry ✓
  - Estimate more **interpretable** metrics: post-editing (PE) effort (human scores, time, % edits to fix)
Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A “hard to beat” baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems ✓
  - MT used in translation industry ✓
  - Estimate more **interpretable** metrics: post-editing (PE) effort (human scores, time, % edits to fix)
  - Some positive results
Some positive results

- **Time to post-edit** subset of sentences predicted as “low PE effort” vs time to post-edit random subset of sentences [Spe11]

<table>
<thead>
<tr>
<th>Language</th>
<th>no QE</th>
<th>QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-en</td>
<td>0.75 words/sec</td>
<td>1.09 words/sec</td>
</tr>
<tr>
<td>en-es</td>
<td>0.32 words/sec</td>
<td>0.57 words/sec</td>
</tr>
</tbody>
</table>

Accuracy in selecting best translation among 4 MT systems [SRT10]

<table>
<thead>
<tr>
<th>Best MT system</th>
<th>Highest QE score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>54%</td>
</tr>
<tr>
<td></td>
<td>77%</td>
</tr>
</tbody>
</table>

Estimating machine translation quality
Some positive results

- **Time to post-edit** subset of sentences predicted as “low PE effort” vs time to post-edit random subset of sentences [Spe11]

<table>
<thead>
<tr>
<th>Language</th>
<th>no QE</th>
<th>QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-en</td>
<td>0.75 words/sec</td>
<td><strong>1.09</strong> words/sec</td>
</tr>
<tr>
<td>en-es</td>
<td>0.32 words/sec</td>
<td><strong>0.57</strong> words/sec</td>
</tr>
</tbody>
</table>
Some positive results

- **Time to post-edit** subset of sentences predicted as “low PE effort” vs time to post-edit random subset of sentences [Spe11]

<table>
<thead>
<tr>
<th>Language</th>
<th>no QE</th>
<th>QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-en</td>
<td>0.75 words/sec</td>
<td><strong>1.09</strong> words/sec</td>
</tr>
<tr>
<td>en-es</td>
<td>0.32 words/sec</td>
<td><strong>0.57</strong> words/sec</td>
</tr>
</tbody>
</table>

- **Accuracy in selecting best translation** among 4 MT systems [SRT10]

<table>
<thead>
<tr>
<th>Best MT system</th>
<th>Highest QE score</th>
</tr>
</thead>
<tbody>
<tr>
<td>54%</td>
<td><strong>77%</strong></td>
</tr>
</tbody>
</table>
Current approaches

Quality indicators

Source text

MT system

Translation

Adequacy indicators

Complexity indicators

Confidence indicators

Fluency indicators

Estimating machine translation quality
Current approaches

- **Quality indicators**
  - Adequacy indicators
  - Complexity indicators
  - Confidence indicators
  - Fluency indicators
  - Source text
  - MT system
  - Translation

- **Learning algorithms**: range of regression, classification, ranking algorithms
Current approaches

- **Quality indicators**
  - Adequacy indicators
  - Complexity indicators
  - Confidence indicators
  - Fluency indicators

- **Learning algorithms**: range of regression, classification, ranking algorithms

- **Datasets**: few with absolute human scores (1-4 scores, PE time, edit distance), WMT data with relative scores
Outline

1. Quality Estimation
2. Shared Task
3. Open issues
4. Conclusions
Objectives

- WMT-12 – joint work with Radu Soricut (Google)
Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on *sentence-level* estimation of PE effort:
Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on sentence-level estimation of PE effort:
  - Identify (new) effective features
Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on sentence-level estimation of PE effort:
  - Identify (new) effective features
  - Identify most suitable machine learning techniques
Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on sentence-level estimation of PE effort:
  - Identify (new) effective features
  - Identify most suitable machine learning techniques
  - Test (new) automatic evaluation metrics
Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on sentence-level estimation of PE effort:
  - Identify (new) effective features
  - Identify most suitable machine learning techniques
  - Test (new) automatic evaluation metrics
  - Establish the state of the art performance in the field
Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on sentence-level estimation of PE effort:
  - Identify (new) effective features
  - Identify most suitable machine learning techniques
  - Test (new) automatic evaluation metrics
  - Establish the state of the art performance in the field
  - Contrast regression and ranking techniques
Objectives
Datasets

English $\rightarrow$ Spanish

- **English** source sentences
Datasets

**English → Spanish**

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
Datasets

**English → Spanish**

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
- **Post-edited** output by 1 professional translator
Datasets

English → Spanish

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
- **Post-edited** output by 1 professional translator
- Effort **scores** by 3 professional translators, scale 1-5, averaged

Human Spanish translation (original references)

# Instances
- Training: 1832
- Blind test: 422
Datasets

**English → Spanish**

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
- **Post-edited** output by 1 professional translator
- Effort **scores** by 3 professional translators, scale 1-5, averaged
- Human Spanish translation (original **references**)
Datasets

English $\rightarrow$ Spanish

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
- **Post-edited** output by 1 professional translator
- Effort **scores** by 3 professional translators, scale 1-5, averaged
- Human Spanish translation (original **references**)
- **# Instances**
  - Training: 1832
  - Blind test: 422
Annotation guidelines

3 human judges for PE effort assigning 1-5 scores for ⟨source, MT output, PE output⟩

[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 to no editing.
Resources provided

SMT resources for training and test sets:
- SMT training corpus (Europarl and News-documentaries)
- LMs: 5-gram LM; 3-gram LM and 1-3-gram counts
- IBM Model 1 table (Giza)
- Word-alignment file as produced by \textit{grow-diag-final}
- Phrase table with word alignment information
- Moses configuration file used for decoding
- Moses run-time log: model component values, word graph, etc.
Two sub-tasks:

- **Scoring**: predict a score in [1-5] for each test instance
- **Ranking**: sort all test instances best-worst
Evaluation metrics

**Scoring metrics** - standard **MAE** and **RMSE**

\[
MAE = \frac{\sum_{i=1}^{N} |H(s_i) - V(s_i)|}{N}
\]

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (H(s_i) - V(s_i))^2}{N}}
\]

\[N = |S|\]

\(H(s_i)\) is the predicted score for \(s_i\)

\(V(s_i)\) the is human score for \(s_i\)
Evaluation metrics

Ranking metrics Spearman’s rank correlation and new metric: DeltaAvg

For $S_1, S_2, \ldots, S_n$ quantiles:

$$\text{DeltaAvg}_V[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S)$$

$V(S)$: extrinsic function measuring the “quality” of set $S$
Evaluation metrics

**Ranking metrics** Spearman’s rank correlation and new metric: DeltaAvg

For $S_1$, $S_2$, $\ldots$, $S_n$ quantiles:

$$\text{DeltaAvg}_V[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n - 1} - V(S)$$

$V(S)$: extrinsic function measuring the “quality” of set $S$

Average human scores (1-5) of set $S$
Evaluation metrics

**DeltaAvg**

Example 1: \( n=2 \), quantiles \( S_1, S_2 \)

\[
\text{DeltaAvg}[2] = V(S_1) - V(S)
\]

“Quality of the top half compared to the overall quality”

Average **human scores** of top half compared to average **human scores** of complete set
Evaluation metrics

![Score Levels]

Average human score: 3
Evaluation metrics

Score categories:
- Score 5
- Score 4
- Score 3
- Score 2
- Score 1

Average human score: 3

Evaluation metrics:
- Random: $[3 - 3] = 0$
- QE: $[3.8 - 3] = 0.8$
- Oracle: $[4.2 - 3] = 1.2$
- Lowerb: $[1.8 - 3] = -1.2$

N = 2

DeltaAvg[2]
Evaluation metrics

Average human score: 3

Random = [3 - 3] = 0
QE = [3.8 - 3] = 0.8
Oracle = [4.2 - 3] = 1.2
Lowerb = [1.8 - 3] = -1.2

Average "human" score of top 50% selected after ranking based on QE score. QE score can be on any scale...
DeltaAvg

Example 2: $n=3$, quantiles $S_1$, $S_2$, $S_3$

\[
\text{DeltaAvg}[3] = \frac{(V(S_1) - V(S)) + (V(S_{1,2}) - V(S))}{2}
\]

Average human scores of top third compared to average human scores of complete set; average human scores of top two thirds compared to average human scores of complete set, averaged
Evaluation metrics

N = 5

DeltaAvg[5]

Random = [3 - 3] = 0
Oracle$_1$ = [5 - 3] = 2
Lowerb$_1$ = [1 - 3] = -2
...
QE$_1$ = [4.1 - 3] = 1.1

Average human score: 3
Evaluation metrics

N = 5
DeltaAvg[5]

Random = [3 - 3] = 0
Oracle₁ = [5 - 3] = 2
Lowerb₁ = [1 - 3] = -2
...
QE₁ = [4.1 - 3] = 1.1
QE₁,₂ = [3.9 - 3] = 0.9

Average human score: 3
Evaluation metrics

Average human score: 3

N = 5

\[ \text{DeltaAvg}[5] = \frac{(1.1 + 0.9 + 0.5 + 0.3)}{4} = 0.7 \]

Random = [3 - 3] = 0
Oracle_1 = [5 - 3] = 2
Lowerb_1 = [1 - 3] = -2

... 

QE_1 = [4.1 - 3] = 1.1
QE_{1,2} = [3.9 - 3] = 0.9
QE_{1,2,3} = [3.5 - 3] = 0.5
QE_{1,2,3,4} = [3.3 - 3] = 0.3
Evaluation metrics

Final DeltaAvg metric

\[ \text{DeltaAvg}_V = \frac{\sum_{n=2}^{N} \text{DeltaAvg}_V[n]}{N - 1} \]

where \( N = |S|/2 \)
Evaluation metrics

**Final DeltaAvg metric**

\[ \text{DeltaAvg}_V = \frac{\sum_{n=2}^{N} \text{DeltaAvg}_V[n]}{N - 1} \]

where \( N = \frac{|S|}{2} \)

Average DeltaAvg\([n]\) for all \( n \), \( 2 \leq n \leq \frac{|S|}{2} \)
## Participants

<table>
<thead>
<tr>
<th>ID</th>
<th>Participating team</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRHLT-UPV</td>
<td>Universitat Politecnica de Valencia, Spain</td>
</tr>
<tr>
<td>UU</td>
<td>Uppsala University, Sweden</td>
</tr>
<tr>
<td>SDLLW</td>
<td>SDL Language Weaver, USA</td>
</tr>
<tr>
<td>Loria</td>
<td>LORIA Institute, France</td>
</tr>
<tr>
<td>UPC</td>
<td>Universitat Politecnica de Catalunya, Spain</td>
</tr>
<tr>
<td>DFKI</td>
<td>DFKI, Germany</td>
</tr>
<tr>
<td>WLV-SHEF</td>
<td>Univ of Wolverhampton &amp; Univ of Sheffield, UK</td>
</tr>
<tr>
<td>SJTU</td>
<td>Shanghai Jiao Tong University, China</td>
</tr>
<tr>
<td>DCU-SYMC</td>
<td>Dublin City University, Ireland &amp; Symantec, Ireland</td>
</tr>
<tr>
<td>UEdin</td>
<td>University of Edinburgh, UK</td>
</tr>
<tr>
<td>TCD</td>
<td>Trinity College Dublin, Ireland</td>
</tr>
</tbody>
</table>

One or two systems per team, most teams submitting for ranking and scoring sub-tasks
Baseline system

Feature extraction software – system-independent features:

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of source 1-grams, 2-grams and 3-grams in frequency quartiles 1 and 4
- % of seen source unigrams
Baseline system

**Feature extraction** software – system-independent features:

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of source 1-grams, 2-grams and 3-grams in frequency quartiles 1 and 4
- % of seen source unigrams

**SVM regression** with RBF kernel with the parameters $\gamma$, $\epsilon$ and $C$ optimized using a grid-search and 5-fold cross validation on the training set.
### Results - ranking sub-task

<table>
<thead>
<tr>
<th>System ID</th>
<th>DeltaAvg</th>
<th>Spearman Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>• SDLLW_M5P_bestDeltaAvg</td>
<td>0.63</td>
<td>0.64</td>
</tr>
<tr>
<td>• SDLLW_SVM</td>
<td>0.61</td>
<td>0.60</td>
</tr>
<tr>
<td>UU_bltk</td>
<td>0.58</td>
<td>0.61</td>
</tr>
<tr>
<td>UU_best</td>
<td>0.56</td>
<td>0.62</td>
</tr>
<tr>
<td>TCD_M5P-resources-only*</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>Baseline (17FFs SVM)</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>PRHLT-UPV</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>UEdin</td>
<td>0.54</td>
<td>0.58</td>
</tr>
<tr>
<td>SJTU</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>WLV-SHEF_FS</td>
<td>0.51</td>
<td>0.52</td>
</tr>
<tr>
<td>WLV-SHEF_BL</td>
<td>0.50</td>
<td>0.49</td>
</tr>
<tr>
<td>DFKI_morphPOSIBM1LM</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>DCU-SYMC_unconstrained</td>
<td>0.44</td>
<td>0.41</td>
</tr>
<tr>
<td>DCU-SYMC_constrained</td>
<td>0.43</td>
<td>0.41</td>
</tr>
<tr>
<td>TCD_M5P-all*</td>
<td>0.42</td>
<td>0.41</td>
</tr>
<tr>
<td>UPC_1</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>UPC_2</td>
<td>0.15</td>
<td>0.19</td>
</tr>
</tbody>
</table>

• = winning submissions  
gray area = not different from baseline  
* = bug-fix was applied after the submission
Results - ranking sub-task

**Oracle methods**: associate various metrics in a oracle manner to the test input:

- **Oracle Effort**: the gold-label Effort
- **Oracle HTER**: the HTER metric against the post-edited translations as reference

<table>
<thead>
<tr>
<th>System ID</th>
<th>DeltaAvg</th>
<th>Spearman Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Effort</td>
<td>0.95</td>
<td>1.00</td>
</tr>
<tr>
<td>Oracle HTER</td>
<td>0.77</td>
<td>0.70</td>
</tr>
<tr>
<td>System ID</td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>SDLLW_M5PbestDeltaAvg</td>
<td>0.61</td>
<td>0.75</td>
</tr>
<tr>
<td>UU_best</td>
<td>0.64</td>
<td>0.79</td>
</tr>
<tr>
<td>SDLLW_SVM</td>
<td>0.64</td>
<td>0.78</td>
</tr>
<tr>
<td>UU_bltk</td>
<td>0.64</td>
<td>0.79</td>
</tr>
<tr>
<td>Loria_SVMlinear</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>UEdin</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>TCD_M5P-resources-only*</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>Baseline (17FFs SVM)</td>
<td>0.69</td>
<td>0.82</td>
</tr>
<tr>
<td>Loria_SVMrbf</td>
<td>0.69</td>
<td>0.83</td>
</tr>
<tr>
<td>SJTU</td>
<td>0.69</td>
<td>0.83</td>
</tr>
<tr>
<td>WLV-SHEF_FS</td>
<td>0.69</td>
<td>0.85</td>
</tr>
<tr>
<td>PRHLT-UPV</td>
<td>0.70</td>
<td>0.85</td>
</tr>
<tr>
<td>WLV-SHEF_BL</td>
<td>0.72</td>
<td>0.86</td>
</tr>
<tr>
<td>DCU-SYMC_unconstrained</td>
<td>0.75</td>
<td>0.97</td>
</tr>
<tr>
<td>DFKI_grcfs-mars</td>
<td>0.82</td>
<td>0.98</td>
</tr>
<tr>
<td>DFKI_cfs-plsreg</td>
<td>0.82</td>
<td>0.99</td>
</tr>
<tr>
<td>UPC_1</td>
<td>0.84</td>
<td>1.01</td>
</tr>
<tr>
<td>DCU-SYMC_constrained</td>
<td>0.86</td>
<td>1.12</td>
</tr>
<tr>
<td>UPC_2</td>
<td>0.87</td>
<td>1.04</td>
</tr>
<tr>
<td>TCD_M5P-all</td>
<td>2.09</td>
<td>2.32</td>
</tr>
</tbody>
</table>
New and effective quality indicators (features)

- Most participating systems use external resources: parsers, POS taggers, NER, etc. $\rightarrow$ variety of features
New and effective quality indicators (features)

- Most participating systems use external resources: parsers, POS taggers, NER, etc. → variety of features
- Many tried to exploit linguistically-oriented features
New and effective quality indicators (features)

- Most participating systems use external resources: parsers, POS taggers, NER, etc. → variety of features
- Many tried to exploit linguistically-oriented features
  - none or modest improvements (e.g. WLV-SHEF)
Discussion

New and effective quality indicators (features)

- Most participating systems use **external resources**: parsers, POS taggers, NER, etc. → variety of features
- Many tried to exploit **linguistically-oriented features**
  - none or modest improvements (e.g. WLV-SHEF)
  - high performance (e.g. “UU” with parse trees)
New and effective quality indicators (features)

- Most participating systems use external resources: parsers, POS taggers, NER, etc. → variety of features
- Many tried to exploit linguistically-oriented features
  - none or modest improvements (e.g. WLV-SHEF)
  - high performance (e.g. “UU” with parse trees)
- Good features:
  - confidence: model components from SMT decoder
New and effective quality indicators (features)

- Most participating systems use external resources: parsers, POS taggers, NER, etc. → variety of features
- Many tried to exploit linguistically-oriented features
  - none or modest improvements (e.g. WLV-SHEF)
  - high performance (e.g. “UU” with parse trees)
- Good features:
  - confidence: model components from SMT decoder
  - pseudo-reference: agreement between 2 SMT systems
  - fuzzy-match like: source (and target) similarity with SMT training corpus (LM, etc)
Discussion

Machine Learning techniques

- Best performing: Regression Trees (M5P) and SVR
Discussion

Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**
  - M5P Regression Trees: compact models, less overfitting, “readable”
Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**
  - M5P Regression Trees: compact models, less overfitting, “readable”
  - SVRs: easily overfit with small training data and large feature set
Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**
  - M5P Regression Trees: compact models, less overfitting, “readable”
  - SVRs: easily overfit with small training data and large feature set
- **Feature selection** crucial in this setup
Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**
  - M5P Regression Trees: compact models, less overfitting, “readable”
  - SVRs: easily overfit with small training data and large feature set
- **Feature selection** crucial in this setup
- **Structured learning** techniques: “UU” submissions (tree kernels)
Discussion

Evaluation metrics

- DeltaAvg → suitable for the ranking task
Discussion

Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task
  - automatic and deterministic (and therefore consistent)
Discussion

Evaluation metrics

- DeltaAvg → suitable for the ranking task
  - automatic and deterministic (and therefore consistent)
  - Extrinsic interpretability
  - Versatile: valuation function $V$ can change, $N$ can change

High correlation with Spearman, but less strict MAE, RMSE → difficult task, values stubbornly high

Regression vs ranking
Most submissions: regression results to infer ranking
Ranking approach is simpler, directly useful in many applications
Discussion

Evaluation metrics

- DeltaAvg → suitable for the ranking task
  - automatic and deterministic (and therefore consistent)
  - Extrinsic interpretability
  - Versatile: valuation function $V$ can change, $N$ can change
  - High correlation with Spearman, but less strict

Estimating machine translation quality
Discussion

Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task
  - automatic and deterministic (and therefore consistent)
  - Extrinsic interpretability
  - Versatile: valuation function $V$ can change, $N$ can change
  - High correlation with Spearman, but less strict

- MAE, RMSE $\rightarrow$ difficult task, values stubbornly high
Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task
  - automatic and deterministic (and therefore consistent)
  - Extrinsic interpretability
  - Versatile: valuation function $V$ can change, $N$ can change
  - High correlation with Spearman, but less strict

- MAE, RMSE $\rightarrow$ difficult task, values stubbornly high

Regression vs ranking

- Most submissions: regression results to infer ranking
Discussion

Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task
  - automatic and deterministic (and therefore consistent)
  - Extrinsic interpretability
  - Versatile: valuation function $V$ can change, $N$ can change
  - High correlation with Spearman, but less strict

- MAE, RMSE $\rightarrow$ difficult task, values stubbornly high

Regression vs ranking

- Most submissions: regression results to infer ranking
- Ranking approach is simpler, directly useful in many applications
Establish state-of-the-art performance

- “Baseline” - hard to beat, previous state-of-the-art
Establish state-of-the-art performance

- “Baseline” - hard to beat, previous state-of-the-art
- Metrics, data sets, and performance points available
Establish state-of-the-art performance

- “Baseline” - hard to beat, previous state-of-the-art
- Metrics, data sets, and performance points available
- Known values for oracle-based upperbounds
- Good resource to further investigate: best features & best algorithms
Follow up

Feature sets available

- 11 systems, 1515 features (some overlap) of various types, from 6 to 497 features per system
- http://www.dcs.shef.ac.uk/~lucia/resources/feature_sets_all_participants.tar.gz
Outline

1. Quality Estimation
2. Shared Task
3. Open issues
4. Conclusions
Agreement between translators

- **Absolute value judgements**: difficult to achieve consistency across annotators even in highly controlled setup
  - 30% of initial dataset discarded: annotators disagreed by more than one category
Agreement between translators

- **Absolute value judgements**: difficult to achieve consistency across annotators even in highly controlled setup
  - 30% of initial dataset discarded: annotators disagreed by more than one category
- **Too subjective?**
More objective ways of generating absolute scores

**TIME**: varies considerably across translators (expected). E.g.: seconds per word
More objective ways of generating absolute scores

**TIME**: varies considerably across translators (expected). E.g.: seconds per word

Can we normalise this variation?
More objective ways of generating absolute scores

**TIME**: varies considerably across translators (expected). E.g.: seconds per word

- Can we normalise this variation?
- A dedicated QE system for each translator?
More objective ways of generating absolute scores

**HTER**: Edit distance between *MT output* and its *minimally post-edited version*
More objective ways of generating absolute scores

**HTER**: Edit distance between MT output and its minimally post-edited version

\[
\text{HTER} = \frac{\#\text{edits}}{\#\text{words\_postedited\_version}}
\]

- Edits: substitute, delete, insert, **shift**

Analysis by Maarit Koponen (WMT-12) on post-edited translations with HTER and 1-5 scores

Translations with low HTER (few edits) & low quality scores (high post-editing effort), and vice-versa

Certain edits seem to require more cognitive effort than others - not captured by HTER
More objective ways of generating absolute scores

**HTER**: Edit distance between **MT output** and its **minimally post-edited version**

\[
HTER = \frac{\#\text{edits}}{\#\text{words\_postedited\_version}}
\]

- Edits: substitute, delete, insert, **shift**
- Analysis by Maarit Koponen (WMT-12) on post-edited translations with HTER and 1-5 scores
  - Translations with low HTER (**few edits**) & low quality scores (**high post-editing effort**), and vice-versa
More objective ways of generating absolute scores

**HTER**: Edit distance between **MT output** and its *minimally post-edited version*

\[
HTER = \frac{\#\text{edits}}{\#\text{words\_postedited\_version}}
\]

- Edits: substitute, delete, insert, **shift**
- Analysis by Maarit Koponen (WMT-12) on post-edited translations with HTER and 1-5 scores
  - Translations with low HTER (**few edits**) & low quality scores (**high post-editing effort**), and vice-versa
  - Certain edits seem to require more cognitive effort than others - not captured by HTER
More objective ways of generating absolute scores

**Keystrokes:** different PE strategies - data from 8 translators (joint work with Maarit Koponen and Wilker Aziz):
More objective ways of generating absolute scores

**Keystrokes:** different PE strategies - data from 8 translators
(joint work with Maarit Koponen and Wilker Aziz):

- Box plots showing seconds per word and HTER for different translators.
- Box plots showing total keystrokes for different translators.
More objective ways of generating absolute scores

PET: http://pers-www.wlv.ac.uk/~in1676/pet/
Use of relative scores

**Ranking of translations**: Suitable if the final application is to compare alternative translations of same source sentence.
Use of relative scores

**Ranking of translations**: Suitable if the final application is to compare alternative translations of same source sentence

- N-best list re-ranking
- System combination
- MT system evaluation
Why do translators use (and trust) TMs?
Why do translators use (and trust) TMs?
Why do translators use (and trust) TMs?

Why can’t we do the same for MT?
Why do translators use (and trust) TMs?

Why can’t we do the same for MT? E.g. Xplanation Group
What is the best metric to estimate PE effort?

- Effort scores are subjective
What is the best metric to estimate PE effort?

- Effort scores are subjective
- Effort/HTER seem to lack “cognitive load”
What is the best metric to estimate PE effort?

- Effort scores are subjective
- Effort/HTER seem to lack “cognitive load”
- Time varies too much across post-editors
What is the best metric to estimate PE effort?

- Effort scores are subjective
- Effort/HTER seem to lack “cognitive load”
- Time varies too much across post-editors
- Keystrokes seem to capture PE strategies, but do not correlate well with PE effort
What is the best metric to estimate PE effort?

- Effort scores are subjective
- Effort/HTER seem to lack “cognitive load”
- Time varies too much across post-editors
- Keystrokes seems to capture PE strategies, but do not correlate well with PE effort
- Source fuzzy match score: as reliable as with TMs?
How to use estimated PE effort scores?

Should (supposedly) bad quality translations be **filtered out** or **shown to translators** (different scores/colour codes as in TMs)?

- Wasting time to read scores and translations vs wasting “gisting” information
How to use estimated PE effort scores?

Should (supposedly) bad quality translations be **filtered out** or **shown to translators** (different scores/colour codes as in TMs)?

- Wasting time to read scores and translations vs wasting “gisting” information
How to use estimated PE effort scores?

How to define a threshold on the estimated translation quality to decide what should be filtered out?

- Translator dependent
- Task dependent
How to use estimated PE effort scores?

How to define a **threshold** on the estimated translation quality to decide what should be filtered out?

- Translator dependent
- Task dependent
How to use estimated PE effort scores?

Do translators prefer detailed estimates (sub-sentence level) or an overall estimate for the complete sentence?

- Too much information vs hard-to-interpret scores
How to use estimated PE effort scores?

Do translators prefer **detailed estimates** (sub-sentence level) or an **overall estimate** for the complete sentence?

- Too much information vs hard-to-interpret scores
- Quality estimation vs error detection
  - IBM’s *Goodness* metric: classifier with sparse binary features (word/phrase pairs, etc.)
Do we really need QE?

Can’t we simply add some good features to SMT models?
Do we really need QE?

Can’t we simply add some good features to SMT models?

- Yes, especially if doing sub-sentence QE/error detection
Do we really need QE?

Can’t we simply add some good features to SMT models?

- Yes, especially if doing sub-sentence QE/error detection
- But not all:
  - Some *linguistically-motivated features* can be difficult/expensive: matching of semantic roles
  - **Global features** are difficult/impossible, e.g: coherence given previous n sentences
Outline

1. Quality Estimation
2. Shared Task
3. Open issues
4. Conclusions
Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort.
It is possible to estimate at least certain aspects of translation quality in terms of PE effort.

PE effort estimates can be used in **real applications**

- Ranking translations: filter out bad quality translations
- Selecting translations from multiple MT systems
Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort
- PE effort estimates can be used in real applications
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- Commercial interest
Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort
- PE effort estimates can be used in real applications
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- Commercial interest
  - SDL LW: TrustScore
  - Multilizer: MT-Qualifier
Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort.
- PE effort estimates can be used in real applications:
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- Commercial interest
  - SDL LW: TrustScore
  - Multilizer: MT-Qualifier
- A number of open issues to be investigated...
Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort.
- PE effort estimates can be used in real applications:
  - Ranking translations: filter out bad quality translations.
  - Selecting translations from multiple MT systems.
- Commercial interest:
  - SDL LW: TrustScore.
  - Multilizer: MT-Qualifier.
- A number of open issues to be investigated...

What we need

Simple, cheap metric like BLEU/fuzzy match level in TMs.
Journal of MT - Special issue

- 15-06-12 - 1st CFP
- 15-08-12 - 2nd CFP
- 5-10-12 - extended submission deadline
- 20-11-12 - reviews due
- January 2013 - camera-ready due (tentative)

WMT-12 QE Shared Task

All feature sets available
Estimating machine translation quality

State-of-the-art systems and open issues

Lucia Specia

University of Sheffield
l.specia@sheffield.ac.uk

6 September 2012
Lucia Specia.
Exploiting Objective Annotations for Measuring Translation Post-editing Effort.

Lucia Specia, Dhwaj Raj, and Marco Turchi.
Machine translation evaluation versus quality estimation.