11-731 Machine Translation

MT Quality Estimation

Alon Lavie
2 April 2015

With Acknowledged Contributions from:
• Lucia Specia (University of Sheffield)
• CCB et al (WMT 2012)
• Radu Soricut et al (SDL Language Weaver)
Outline

- Quality Estimation Measures:
  - What are they and why they are needed?
- Applications
- Framework and Types of Features
- The WMT 2012 Shared Task on Quality Estimation
- Case Study: The SDL/Language Weaver QE System for WMT 2012
- Open Issues
- Conclusions
MT Quality Estimation

- MT Systems are used in a variety of applications and scenarios
  - Need to assess how well they are performing and whether they are suitable for the task in which they are being used
  - MT systems perform best on input similar to their training data
  - System performance can vary widely from one sentence to the next
- MT Evaluation metrics can provide offline information:
  - Pre-selected test data with human reference translation to compare against
  - Metrics: BLEU, Meteor, TER
- What about online assessment in real time?
  - No human reference translation
  - Needs to be computable in real-time
MT Quality Estimation

Main Driving Applications:
- Is an MT-translated document sufficient in quality for publication and/or user consumption?
  - Example: Translated product reviews or recommendations – publish?
  - Example: Translated news summaries - sufficient for gisting?

MT translation used as a first-step for human translation:
- Pre-translate a document with MT or use Translation Memory?
- Is an MT-generated translation segment worth post-editing? Faster and better than translating the segment from scratch?
- Should poor quality MT-generated segments be filtered out?
- Can we predict in advance how much time/effort will it take to post-edit a document?

Hypothesis Selection and MT System Combination:
- Select the better output from multiple systems
MT Quality Estimation: Framework

- Supervised Learning Task
  - Learn from examples of MT-generated translations and human-generated quality assessments to predict assessments for new unseen MT-generated translation outputs

- What level of granularity?
  - Document-level or segment-level?

- What types of assessments?
  - Quality scale based on human judgments
    - Adequacy/Fluency [1-5] [0-1]
    - Post-editing effort [1-4] [0-1]
    - Class labels: Bad/OK/Good

- What type of machine learning?
  - Classifiers for two or more classes [Good/Bad] [Good/OK/Bad]
  - Logistic regression to maximize correlation with human label scales
  - Ranking algorithms to maximize ranking correlation with human data
MT Quality Estimation: Framework

- What types of features?
  - No reference translation available!
  - Indicators extracted from the **MT-generated output** itself
    - Output length, lexical features, linguistic complexity, LM-based
  - Indicators extracted from the **source-language input**
    - Input length, lexical features, linguistic complexity, LM-based
  - Indicators extracted from **MT system internal features**
    - Decoder features scores: translation model, LM, rules applied
  - **Other features**
    - OOV words, source-target similarity, similarity to training data
    - Deeper linguistic analysis features
MT Quality Estimation: Framework

Quality Estimation Indicators:

- Adequacy indicators
- Complexity indicators
- Confidence indicators
- Fluency indicators
MT Quality Estimation: Framework

- **Training:**

- **Runtime:**
MT Quality Estimation: History

- Similar ideas in the context of MT System Combination around from the 1990s
- Some preliminary exploration in the form of “Confidence Estimation” in 2001/2002 inspired by confidence scores in speech recognition (word posterior probabilities)
- JHU Summer Workshop 2003:
  - Goal: Predict BLEU/NIST/WER scores at runtime
  - Relatively weak MT systems at the time
  - Poor results
- New surge of interest since 2008:
  - Better MT systems
  - MT increasingly used for post-editing
  - More meaningful human scores as data: post-editing time/effort
Some Recent 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>54%</td>
<td>77%</td>
</tr>
</tbody>
</table>
WMT 2012 QE Shared Task

- First large-scale competitive shared-task on Quality Estimation systems:
  - Coordinated by Lucia Specia and Radu Soricut at WMT 2012
  - Provide a common setting for development and comparison of QE systems
  - Focus on **sentence-level QE of Post-Editing Effort**

- Main Objectives:
  - Identify (new) effective features
  - Identify most suitable machine learning techniques
  - Contrast regression and ranking techniques
  - Test (new) automatic evaluation metrics
  - Establish the state of the art performance on this problem
WMT 2012 QE Shared Task

Data and Setting:
- Single common MT system generating data
- English to Spanish
- Moses Phrase-based SMT system developed on WMT 2012 data
- English source sentences; Spanish MT-generated output sentences
- MT output post-edited by a single professional translator
- Post-editing effort scored by three independent translators using a discrete [1-5] scale; averaged for each segment
- Spanish human reference translations available for analysis but not disclosed to QE development teams
- Data made available for development: 1832 segments
- Blind (unseen) test data: 422 segments
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.
WMT 2012 QE Shared Task

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 *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.
WMT 2012 QE Shared Task

- Two Sub-tasks:
  - **Scoring**: Predict a post-editing effort score [1-5] for each test segment
  - **Ranking**: Rank the test segments from best to worst
Scoring Task evaluation measures:
- Mean-Absolute-Error (MAE)
- Root-Mean-Squared-Error (RMSE)

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

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

\(N = |S|\)

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

\(V(s_i)\) the is human score for \(s_i\)
WMT 2012 QE Shared Task

- **Ranking Task** evaluation measures:
  - Spearman’s Rank Correlation Coefficient
    \[ \rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}. \]
  - 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 \( S \)
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
Average human score: 3

\begin{align*}
\text{Random} & = [3, 3] = 0 \\
\text{QE} & = [3.8, 3] = 0.8 \\
\text{Oracle} & = [4.2 - 3] = 1.2 \\
\text{Lowerb} & = [1.8 - 3] = -1.2
\end{align*}

Average “human” score of top 50% selected after ranking based on QE score. QE score can be on any scale...
Final DeltaAvg metric

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

where $N = |S|/2$

Average DeltaAvg$[n]$ for all $n$, $2 \leq n \leq |S|/2$
WMT 2012 QE Shared Task

- **Participating Teams:**

<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 Features and 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
## WMT 2012 QE Shared Task

### Results - Ranking Task:

<table>
<thead>
<tr>
<th>System ID</th>
<th>DeltaAvg</th>
<th>Spearman Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDLLW_M5PbestDeltaAvg</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
WMT 2012 QE Shared Task

- Ranking Task - Oracles:

**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>
</tbody>
</table>
# WMT 2012 QE Shared Task

## Results - Scoring Task:

<table>
<thead>
<tr>
<th>System ID</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<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_SVMrbuf</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>
Analysis:

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)
WMT 2012 QE Shared Task

- Analysis:

  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)
WMT 2012 QE Shared Task

Analysis:

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
WMT 2012 QE Shared Task

- Analysis:

  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
Case Study: SDL LW QE System

- Best performing system(s) in WMT 2012 shared tasks
- Two main system variants:
  - M5P Regression Tree model
  - SVM Regression Model (SVR)
- Main distinguishing characteristics:
  - Novel Features used
  - Feature Selection was crucial to performance
  - Machine Learning approaches used
Case Study: SDL LW QE System

- **Features Used:**
  - Total number of features: 42
  - Baseline Features: 17
  - Decoder Features: 8
  - New LW Features: 17
Case Study: SDL LW QE System

● Baseline Features:

BF1 number of tokens in the source sentence
BF2 number of tokens in the target sentence
BF3 average source token length
BF4 LM probability of source sentence
BF5 LM probability of the target sentence
BF6 average number of occurrences of the target word within the target translation
BF7 average number of translations per source word in the sentence (as given by IBM 1 table thresholded so that $Prob(t|s) > 0.2$)
BF8 average number of translations per source word in the sentence (as given by IBM 1 table thresholded so that $Prob(t|s) > 0.01$) weighted by the inverse frequency of each word in the source corpus
BF9 percentage of unigrams in quartile 1 of frequency (lower frequency words) in SMT$_{src}$
BF10 percentage of unigrams in quartile 4 of frequency (higher frequency words) in SMT$_{src}$
BF11 percentage of bigrams in quartile 1 of frequency of source words in SMT$_{src}$
BF12 percentage of bigrams in quartile 4 of frequency of source words in SMT$_{src}$
BF13 percentage of trigrams in quartile 1 of frequency of source words in SMT$_{src}$
BF14 percentage of trigrams in quartile 4 of frequency of source words in SMT$_{src}$
BF15 percentage of unigrams in the source sentence seen in SMT$_{src}$
BF16 number of punctuation marks in source sentence
BF17 number of punctuation marks in target sentence
Case Study: SDL LW QE System

- Moses-based Decoder Features:
  
  MF1  Distortion cost
  
  MF2  Word penalty cost
  
  MF3  Language-model cost
  
  MF4  Cost of the phrase-probability of source given target $\Phi(s|t)$
  
  MF5  Cost of the word-probability of source given target $\Phi_{lex}(s|t)$
  
  MF6  Cost of the phrase-probability of target given source $\Phi(t|s)$
  
  MF7  Cost of the word-probability of target given source $\Phi_{lex}(t|s)$
  
  MF8  Phrase penalty cost
Case Study: SDL LW QE System

- New LW Features:

- LF1 number of out-of-vocabulary tokens in the source sentence
- LF2 LM perplexity for the source sentence
- LF3 LM perplexity for the target sentence
- LF4 geometric mean ($\lambda$-smoothed) of 1-to-4-gram precision scores (i.e., BLEU score without brevity-penalty) of source sentence against the sentences of $SMT_{src}$ used as “references”
- LF5 geometric mean ($\lambda$-smoothed) of 1-to-4-gram precision scores of target translation against the sentences of $SMT_{trg}$ used as “references”
- LF6 geometric mean ($\lambda$-smoothed) of 1-to-4-gram precision scores of source sentence against the top BLEU-scoring quartile of $Dev_{src}$
- LF7 geometric mean ($\lambda$-smoothed) of 1-to-4-gram precision scores of target translation against the top BLEU-scoring quartile of $Dev_{trg}$
- LF8 geometric mean ($\lambda$-smoothed) of 1-to-4-gram precision scores of source sentence against the bottom BLEU-scoring quartile of $Dev_{src}$
- LF9 geometric mean ($\lambda$-smoothed) of 1-to-4-gram precision scores of target translation against the bottom BLEU-scoring quartile $Dev_{trg}$
- LF10 geometric mean ($\lambda$-smoothed) of 1-to-4-gram precision scores of target translation against a pseudo-reference produced by a second MT Eng-Spa system
- LF11 count of one-to-one (O2O) word alignments between source and target translation
- LF12 ratio of O2O alignments over source sentence
- LF13 ratio of O2O alignments over target translation
- LF14 count of O2O alignments with Part-of-Speech-agreement
- LF15 ratio of O2O alignments with Part-of-Speech-agreement over O2O alignments
- LF16 ratio of O2O alignments with Part-of-Speech-agreement over source
- LF17 ratio of O2O alignments with Part-of-Speech-agreement over target
## Case Study: SDL LW QE System

- **Results – Baseline Features:**

<table>
<thead>
<tr>
<th>Systems</th>
<th>Ranking</th>
<th>Scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DeltaAvg</td>
<td>Spearman</td>
</tr>
<tr>
<td>17 BF with M5P</td>
<td>0.53</td>
<td>0.56</td>
</tr>
<tr>
<td>17 BF with SVR</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>best-system</td>
<td>0.63</td>
<td>0.64</td>
</tr>
</tbody>
</table>

*Table 1: Performance of the Baseline Features using M5P and SVR models on the test set.*
Case Study: SDL LW QE System

- Results - Moses-based Features:

<table>
<thead>
<tr>
<th>Systems</th>
<th>Ranking DeltaAvg</th>
<th>Ranking Spearman-Corr</th>
<th>MAE</th>
<th>RMSE</th>
<th>Predict. Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 MFs with M5P</td>
<td>0.58</td>
<td>0.58</td>
<td>0.65</td>
<td>0.81</td>
<td>[1.8-5.0]</td>
</tr>
<tr>
<td>best-system</td>
<td>0.63</td>
<td>0.64</td>
<td>0.61</td>
<td>0.75</td>
<td>[1.7-5.0]</td>
</tr>
</tbody>
</table>

Table 2: Performance of the Moses-based Features with an M5P model on the test set.
Case Study: SDL LW QE System

- Results – All Features:

<table>
<thead>
<tr>
<th>Systems</th>
<th>#L.Eq</th>
<th>Dev Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DeltaAvg</td>
<td>MAE</td>
</tr>
<tr>
<td>42 FFs with M5P</td>
<td>10</td>
<td>0.60</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>(best-system)</strong> 15 FFs with M5P</td>
<td>2</td>
<td><strong>0.63</strong></td>
<td><strong>0.52</strong></td>
</tr>
<tr>
<td>14 FFs with M5P</td>
<td>6</td>
<td>0.62</td>
<td><strong>0.50</strong></td>
</tr>
</tbody>
</table>

Table 3: M5P-model performance for different feature-function sets (15-FFs ∈ 42-FFs; 14-FFs ∈ 42-FFs).
Case Study: SDL LW QE System

- Best Features:

<table>
<thead>
<tr>
<th>DeltaAvg optim.</th>
<th>BF1  BF3  BF4  BF6  BF12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BF13 BF14 MF3 MF4 MF6</td>
</tr>
<tr>
<td></td>
<td>LF1 LF10 LF14 LF15 LF16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAE optim.</th>
<th>BF1  BF3  BF4  BF6  BF12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BF14 BF16 MF3 MF4 MF6</td>
</tr>
<tr>
<td></td>
<td>LF1 LF10 LF14 LF17</td>
</tr>
</tbody>
</table>

Table 4: Feature selection results.
Case Study: SDL LW QE System

- **Best Features (MAE Optimal):**
  - BF1: number of tokens in the source sentence
  - BF3: average source token length
  - BF4: LM probability of source sentence
  - BF6: average number of occurrences of the target word within the target translation
  - BF12: percentage of bigrams in quartile 4 of frequency of source words in $SMT_{src}$
  - BF14: percentage of trigrams in quartile 4 of frequency of source words in $SMT_{src}$
  - BF16: number of punctuation marks in source sentence
  - MF3: Language Model cost
  - MF4: cost of the phrase-probability of source given target
  - MF6: cost of the phrase-probability of target given source
  - LF1: number of out-of-vocabulary tokens in the source sentence
  - LF10: geometric mean of 1-to-4-gram precision scores of target translation against a pseudo-reference produced by a second EN-to-ES MT system
  - LF14: count of 1-to-1 alignments with Part-of-Speech-agreement
  - LF17: ratio of 1-to-1 alignments with Part-of-Speech-agreement over target
Open Issues

- Agreement between Translators:
  - Noisy “Gold standard” PE effort data
    - Absolute value judgments: difficult to achieve consistency across annotators even in highly controlled setup
    - 30% of initial dataset discarded: annotators disagreed by more than one category
  - Need for better methodology in establishing PE effort
  - HTER is not a great solution:
    - **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
Open Issues

- How to utilize QE scores as estimated post-editing effort scores?
  - Should (supposedly) bad quality translations be filtered out or shown to translators (different scores/color codes)?
  - Tradeoff of translator wasting time looking at MT segments with bad scores/colors versus translators missing out on useful information

- How to define a threshold on the estimated translation quality to decide which MT segments should be filtered out?
  - Translator dependent?
  - Task dependent?
    - Output quality and project time requirements

- Should the focus instead be on identifying the likely errors in the MT output rather than on estimating how good it is?
Open Issues

Do we really need QE? Can’t we use these features to directly improve or correct the MT output?

- In some cases yes, based on sub-sentence QE/error detection
- Generally, this is very difficult:
  - Some linguistically-motivated features can be difficult and expensive to integrate into decoding (e.g. matching of semantic roles)
  - Global features are particularly very difficult to incorporate into decoding, (e.g: coherence given previous n sentences)

Michael Denkowski’s PhD thesis addresses many of these issues:

- Immediate incremental learning of translation models from translator post-edited segments
- Tuning of features to learn how much to trust such incremental information
- New advanced MT evaluation metrics that directly reflect post-editing effort - optimizing MT systems to such metrics
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
- Significant and growing commercial interest in this problem
- Challenging research problem with lots of open issues and questions to work on!