Human Evaluation and Human Parity for Chinese-English News Translation

Christian Federmann
Microsoft Translator
Microsoft Translator and MSRA Teams

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou

arxiv paper: https://arxiv.org/abs/1803.05567
“First step on the trajectory towards human parity for machine translation”
Project Babel: a roadmap to Human Parity

- Define new challenge for NMT research
  - MT quality has improved a lot:
    - How far are we from human performance?
    - Fundamental question: How can we measure this?

- 2016 – Near Parity
  - The Verge: In some cases, Google says its GNMT system is even approaching human-level translation accuracy. That near-parity is restricted to transitions between related languages, like from English to Spanish and French.

- 2018 – Human Parity
  - Microsoft researchers achieve human parity for distant language pair Chinese to English
Outrageous!

Focus on
Human evaluation

Goal
Measure human parity
Defining Human Parity

**Direct, equivalence-based definition**

If a bilingual human judges the quality of a candidate translation produced by a human to be equivalent to one produced by a machine, then the machine has achieved human parity.

But... hard to determine “equivalence” of translation quality
Defining Human Parity

Indirect, difference-based definition

If there is no statistically significant difference between human quality scores for a test set of candidate translations from a machine translation system and the scores for the corresponding human translations then the machine has achieved human parity.

Given a reliable scoring metric, we can measure this!
Defining Human Parity

From

(Human == Machine) \rightarrow \text{Human Parity}

To

\neg (\text{Human} <> \text{Machine}) \rightarrow \text{Human Parity}
Defining Human Parity

Assumptions

1. Possible to measure MT quality using sampled test sets
2. Possible to measure MT quality using aggregated segment scores
3. Reliable scoring metric exists

Notes

• No claim of superiority!
• Translation not necessarily error-free
• Results valid on chosen test set only
Why not use BLEU?

**Automatic metrics**
- Use BLEU with high quality references?
- Quality issues with original WMT reference

- Created two new references:
  - PE = post-edited / crowd-sourced
  - HT = human translation from scratch

**Reference bias**
- Online-B-1710?

**Conclusions**
- There is no “human BLEU score”
- Use source-based, human evaluation

![BLEU scores against HT, PE, WMT references](chart)
Measuring Human Parity

Requirements

• Reliable scoring metric: direct assessment (DA), following state-of-the-art WMT17
• Modified to use source-based evaluation, following IWSLT17
• Enforced full overlap for all systems, with triple annotator redundancy per segment

Evaluation design

• Regular evaluation campaigns over time (difference to WMT evals, which are static)
• Final evaluation campaign based on 3x Subset-1, Subset-2, Subset-3, and Subset-4
• Collected similar amount of annotations as for WMT17 → large-scale, reliable eval!
• Covering nearly half of the WMT17 test set
Direct assessment

Simple task
- Assigns absolute score relative to "translation hint"
- In our case, relative to source text
- Each task contains 100 items

Reliable scores
- Embedded quality control data
- Monitor annotator reliability
- Enforced segment overlap
Campaigns

Timeline
- Regular monthly evaluation campaigns
- Final round of evaluation campaigns in February/March 2018

Scale
- 9 systems under investigation, including 3 research candidates
- 15 annotators per subset, 6 subsets
- 20 tasks per subset, 3 redundant annotators per task
- 4,200 data points per subset (excluding quality controls)
- 25,200 data points across all subsets
# Project Babel: Monthly evaluations

<table>
<thead>
<tr>
<th>November 2017</th>
<th>Rank group</th>
<th>December 2017</th>
<th>Rank group</th>
<th>January 2018</th>
<th>Rank group</th>
<th>February 2018</th>
<th>Rank group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pactera human translation</td>
<td>1</td>
<td>Unbabel post-edited</td>
<td>1</td>
<td>Unbabel post-edited</td>
<td>1</td>
<td>Patera human trans</td>
<td>1</td>
</tr>
<tr>
<td>Unbabel post-edited</td>
<td>1</td>
<td>Pactera human translation</td>
<td>1</td>
<td>Pactera human trans</td>
<td>2</td>
<td>Unbabel post-edited</td>
<td>1</td>
</tr>
<tr>
<td>Sogou Knowing NMT</td>
<td>2</td>
<td><strong>Dual Learning TF/Transformer</strong></td>
<td>2</td>
<td><strong>MSR Redmond 20180112</strong></td>
<td>3</td>
<td><strong>MSRA ML 20180211</strong></td>
<td>3</td>
</tr>
<tr>
<td><em>MSR 20171012 research</em></td>
<td>3</td>
<td><strong>Transformer+R2L</strong></td>
<td>2</td>
<td><strong>MSRA NLC 20180108</strong></td>
<td>3</td>
<td><strong>MSRA ML 20180211</strong></td>
<td>3</td>
</tr>
<tr>
<td>Online-A-1710</td>
<td>4</td>
<td><strong>Karnak/Transformer</strong></td>
<td>2</td>
<td>Sogou Knowing NMT</td>
<td>4</td>
<td>Sogou Knowing NMT</td>
<td>4</td>
</tr>
<tr>
<td>Online-B-1710</td>
<td>4</td>
<td><strong>Dual Learning Karnak/NMT</strong></td>
<td>3</td>
<td><strong>MSRA ML 20180111</strong></td>
<td>4</td>
<td>newstest 2017 reference</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>TF/Transformer</strong></td>
<td>4</td>
<td>Online-A-1710</td>
<td>4</td>
<td>Online-A-1710</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Transformer+R2L+BackTrans</strong></td>
<td>4</td>
<td>Online-B-1710</td>
<td>5</td>
<td>Online-B-1710</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Online-A-1710</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Evaluation against fixed-points: Two human translations, online MT systems, and Sogou Sogou is winner of WMT2017 competition Evaluating various research systems (*italics*)
Result clusters

Tabular representation

• Clustering boils down to pairwise differences and their significance
• Clusters based on number of significant wins against all lower ranked systems
• Systems within same cluster are considered indistinguishable
• Wilcoxon rank sum test

<table>
<thead>
<tr>
<th>Rank</th>
<th>Z score</th>
<th>R score</th>
<th>System ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.237</td>
<td>69.0</td>
<td>Combo-6</td>
</tr>
<tr>
<td></td>
<td>0.220</td>
<td>68.5</td>
<td>Reference-HT</td>
</tr>
<tr>
<td></td>
<td>0.216</td>
<td>68.9</td>
<td>Combo-5</td>
</tr>
<tr>
<td></td>
<td>0.211</td>
<td>68.6</td>
<td>Combo-4</td>
</tr>
<tr>
<td></td>
<td>0.141</td>
<td>67.3</td>
<td>Reference-PE</td>
</tr>
<tr>
<td>2</td>
<td>-0.094</td>
<td>62.3</td>
<td>Sogou</td>
</tr>
<tr>
<td></td>
<td>-0.115</td>
<td>62.1</td>
<td>Reference-WMT</td>
</tr>
<tr>
<td>3</td>
<td>-0.398</td>
<td>56.0</td>
<td>Online-A-1710</td>
</tr>
<tr>
<td></td>
<td>-0.468</td>
<td>54.1</td>
<td>Online-B-1710</td>
</tr>
</tbody>
</table>
Visualising Human Parity

From

Result clusters for all systems

To

Pairwise density representation
Combo-6 vs Sogou

Score distributions for zho to eng in BabelEval5_2_ALL

- Combo-6 density (n=1629)
- Combo-6 mean (68.96)
- Sogou density (n=1629)
- Sogou mean (62.33)
Combo-6 vs WMT

Score distributions for zho to eng in BabelEval5_2_ALL

- Combo-6 density (n=1629)
- Combo-6 mean (68.96)
- WMT density (n=1629)
- WMT mean (62.06)
Score distributions for zho to eng in BabelEval5_2_ALL

- Combo-6 density (n=1629)
- Combo-6 mean (68.96)
- PE density (n=1629)
- PE mean (67.26)
Score distributions for zho to eng in BabelEval5_2_ALL

- Combo-6 density (n=1629)
- Combo-6 mean (68.96)
- HT density (n=1629)
- HT mean (68.55)
Combo-6 vs HT

Score distributions for zho to eng in BabelEval5_2_ALL

- Combo-6 density (n=1629)
- Combo-6 mean (68.96)
- HT density (n=1629)
- HT mean (68.55)

Human Parity!
Where do we go from here?

Open data

- Released all data, including new reference translations → fostering future research
- [https://github.com/MicrosoftTranslator/Translator-HumanParityData](https://github.com/MicrosoftTranslator/Translator-HumanParityData)

Improved quality

- Extend human parity to consider contextual information
- Measure quality against human certification levels

Challenging future

- First step on trajectory towards human parity for machine translation
- New languages, domains, architectures