MT Eval Workshop LREC’02

Report by Eva Forsbom,
evafo@stp.ling.uu.se,
Uppsala University,
Department of Linguistics
Human-Based Metric

Correctness/Adequacy/Fidelity

In this evaluation, a human evaluator grades the adequacy of translated segments as compared to a reference translation according to the following scale:

1. None of the meaning expressed in the source fragment is expressed in the translation fragment
2. Little of the source fragment meaning is expressed in the translation fragment
3. Much of the source fragment meaning is expressed in the translation fragment
4. Most of the source fragment meaning is expressed in the translation fragment
5. All meaning expressed in the source fragment appears in the translation fragment
Human-Based Metric

Strong Point

- Inter-evaluator consistency – similar rankings.

Weak Points

- Intra-evaluator consistency – drifting grading level when several translations of the same text.
- Inter-evaluator consistency – different grading levels.
- Interference/Compensation – syntax, stylistics, and spelling affect the grading, by over- or undercompensated grades.
- Quality of reference translation – “verbatim” reference translations tend to penalise freer, often better, translations.
- Segment length – longer segments tend to get lower grades.
- Alternative translations – non-disambiguated translations tend to get higher grades.
# Human-Based Metric

## Results – Score by Evaluator, 1

<table>
<thead>
<tr>
<th>Transl</th>
<th>Eval1</th>
<th>Eval2</th>
<th>Eval3</th>
<th>Eval4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>57.5</td>
<td>58.8</td>
<td>41.2</td>
<td>60.0</td>
<td>41.4</td>
</tr>
<tr>
<td>102</td>
<td>48.8</td>
<td>55.0</td>
<td>38.8</td>
<td>55.0</td>
<td>37.6</td>
</tr>
<tr>
<td>103</td>
<td>47.5</td>
<td>56.2</td>
<td>35.0</td>
<td>61.3</td>
<td>38.1</td>
</tr>
<tr>
<td>104</td>
<td>55.0</td>
<td>57.5</td>
<td>43.8</td>
<td>61.3</td>
<td>41.4</td>
</tr>
<tr>
<td>105</td>
<td>52.5</td>
<td>53.8</td>
<td>31.2</td>
<td>56.2</td>
<td>36.9</td>
</tr>
<tr>
<td>106</td>
<td>52.5</td>
<td>56.2</td>
<td>32.5</td>
<td>58.8</td>
<td>38.1</td>
</tr>
<tr>
<td>107</td>
<td>35.0</td>
<td>40.0</td>
<td>25.0</td>
<td>61.3</td>
<td>30.7</td>
</tr>
<tr>
<td>108</td>
<td>46.2</td>
<td>52.5</td>
<td>28.7</td>
<td>60.0</td>
<td>35.7</td>
</tr>
<tr>
<td>109</td>
<td>47.5</td>
<td>55.0</td>
<td>28.7</td>
<td>61.3</td>
<td>36.7</td>
</tr>
<tr>
<td>110</td>
<td>42.5</td>
<td>53.8</td>
<td>32.5</td>
<td>61.3</td>
<td>36.2</td>
</tr>
<tr>
<td>111</td>
<td>40.0</td>
<td>55.0</td>
<td>30.0</td>
<td>61.3</td>
<td>35.5</td>
</tr>
<tr>
<td>112</td>
<td>37.5</td>
<td>52.5</td>
<td>30.0</td>
<td>58.8</td>
<td>34.0</td>
</tr>
<tr>
<td>113</td>
<td>35.0</td>
<td>55.0</td>
<td>35.0</td>
<td>61.3</td>
<td>35.5</td>
</tr>
</tbody>
</table>
## Human-Based Metric

### Results – Score by Evaluator, 2

<table>
<thead>
<tr>
<th>Transl</th>
<th>Eval1</th>
<th>Eval2</th>
<th>Eval3</th>
<th>Eval4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>23.8</td>
<td>27.6</td>
<td>18.1</td>
<td>24.8</td>
<td>30.9</td>
</tr>
<tr>
<td>202</td>
<td>25.7</td>
<td>33.3</td>
<td>21.9</td>
<td>35.2</td>
<td>38.1</td>
</tr>
<tr>
<td>203</td>
<td>32.4</td>
<td>37.1</td>
<td>31.4</td>
<td>37.1</td>
<td>45.3</td>
</tr>
<tr>
<td>204</td>
<td>28.6</td>
<td>34.3</td>
<td>22.9</td>
<td>36.2</td>
<td>40.0</td>
</tr>
<tr>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>207</td>
<td>15.2</td>
<td>21.0</td>
<td>14.3</td>
<td>27.6</td>
<td>25.6</td>
</tr>
<tr>
<td>208</td>
<td>18.1</td>
<td>33.3</td>
<td>17.1</td>
<td>35.2</td>
<td>34.1</td>
</tr>
<tr>
<td>209</td>
<td>18.1</td>
<td>30.5</td>
<td>16.2</td>
<td>33.3</td>
<td>32.2</td>
</tr>
<tr>
<td>210</td>
<td>23.8</td>
<td>32.4</td>
<td>18.1</td>
<td>35.2</td>
<td>35.9</td>
</tr>
<tr>
<td>211</td>
<td>20.0</td>
<td>34.3</td>
<td>18.1</td>
<td>35.2</td>
<td>35.3</td>
</tr>
<tr>
<td>212</td>
<td>16.2</td>
<td>24.8</td>
<td>14.3</td>
<td>32.4</td>
<td>28.7</td>
</tr>
<tr>
<td>213</td>
<td>22.9</td>
<td>32.4</td>
<td>16.2</td>
<td>35.2</td>
<td>35.0</td>
</tr>
</tbody>
</table>
Human-Based Metric

Results – Rank by Evaluator, 1

<table>
<thead>
<tr>
<th>Transl</th>
<th>Eval1</th>
<th>Eval2</th>
<th>Eval3</th>
<th>Eval4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>102</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>103</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>104</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>3</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>106</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>107</td>
<td>12</td>
<td>13</td>
<td>13</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>108</td>
<td>8</td>
<td>11</td>
<td>11</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>109</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>110</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>111</td>
<td>10</td>
<td>5</td>
<td>9</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>112</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>113</td>
<td>12</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>
Human-Based Metric

Results – Rank by Evaluator, 2

<table>
<thead>
<tr>
<th>Transl</th>
<th>Eval1</th>
<th>Eval2</th>
<th>Eval3</th>
<th>Eval4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>202</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>203</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>204</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>207</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>208</td>
<td>8</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>209</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>210</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>211</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>212</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>213</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>
Automated Metric

Named Entity Translations

In this evaluation, some human annotators marks up named entities (NE) in a reference translation. All unique NE’s from the reference translation are then searched in the translations, and all unique occurrences counted. Some normalisation processes could also be applied:

- No normalisation (NONE)
- Case folding (CASE)
- Diacritica to non-diacritica conversion (DIA)
- Number normalisation (NUMB)
- Removal of possesives (no occurrence)
- Combinations (CASE&DIA, ..., ALL)
Automated Metric

Strong Point

- Measures NE translations.

Weak Points

- Only relevant when many NE’s.
- Diacritica conversion tends to boost bad translations (e.g. Émirate)
- Quality of reference translation matters (e.g. 60 %).
- Does not seem to correlate with the adequacy metric, not even before normalisation.
## Automated Metric

### Results – Score by Norm. Type, 1

<table>
<thead>
<tr>
<th>Transl</th>
<th>NONE</th>
<th>NUMB</th>
<th>DIA</th>
<th>CASE&amp;DIA</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>101</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>102</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>103</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>104</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>105</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>106</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>107</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>108</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>109</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>110</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>111</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>112</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>113</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
# Automated Metric

## Results – Score By Norm. Type, 2

<table>
<thead>
<tr>
<th>Transl</th>
<th>NONE</th>
<th>NUMB</th>
<th>DIA</th>
<th>CASE&amp;DIA</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>201</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>202</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>203</td>
<td>87.5</td>
<td>87.5</td>
<td>87.5</td>
<td>87.5</td>
<td>87.5</td>
</tr>
<tr>
<td>204</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>207</td>
<td>75.0</td>
<td>87.5</td>
<td>75.0</td>
<td>87.5</td>
<td>100.0</td>
</tr>
<tr>
<td>208</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
</tr>
<tr>
<td>209</td>
<td>50.0</td>
<td>50.0</td>
<td>62.5</td>
<td>50.0</td>
<td>62.5</td>
</tr>
<tr>
<td>210</td>
<td>62.5</td>
<td>75.0</td>
<td>62.5</td>
<td>62.5</td>
<td>75.0</td>
</tr>
<tr>
<td>211</td>
<td>62.5</td>
<td>75.0</td>
<td>62.5</td>
<td>62.5</td>
<td>75.0</td>
</tr>
<tr>
<td>212</td>
<td>87.5</td>
<td>100.0</td>
<td>87.5</td>
<td>87.5</td>
<td>100.0</td>
</tr>
<tr>
<td>213</td>
<td>62.5</td>
<td>75.0</td>
<td>62.5</td>
<td>62.5</td>
<td>75.0</td>
</tr>
</tbody>
</table>
## Automated Metric

### Results – Rank by Norm. Type, 1

<table>
<thead>
<tr>
<th>Transl</th>
<th>NONE</th>
<th>NUMB</th>
<th>DIA</th>
<th>CASE&amp;DIA</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>102</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>103</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>104</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>106</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>107</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>108</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>109</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>110</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>111</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>112</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>113</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Automated Metric

Results – Rank By Norm. Type, 2

<table>
<thead>
<tr>
<th>Transl</th>
<th>NONE</th>
<th>NUMB</th>
<th>DIA</th>
<th>CASE&amp;DIA</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>202</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>203</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>204</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>207</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>208</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>209</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>210</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>211</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>212</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>213</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>