Human-centered Analysis of Machine Translation Quality

Aljoscha Burchardt (DFKI)

Joint work with Eleftherios Avramidis, António Branco, Kim Harris, Arle Lommel, Vivien Macketanz, Lucia Specia, Marco Turchi, Hans Uszkoreit, and others
Assessing quality in MT development

MT Engine

multi-bleu.perl

?
Why?

- Current statistical machine translation has its roots in gisting translation (aka information translation)
- Goal: Improvement on average


MT Evaluation is Difficult

• In many NLP tasks, performance can be measured as deviation from some ideal (POS tagging, parsing, fact extraction, etc.)

• In MT, this is difficult
  • Theoretical issue: there is no eternal notion of “good translation”, MT quality is task-specific.
  • Practical issue: there are usually many different good translations, no simple notion of deviation.

• Example:
  • Input: Use your antivirus to perform a complete scanning.
  • MT output: Verwenden Sie Ihre Antivirus eine vollständige Abtastung durchzuführen.
  • Translator 1: Benutzen Sie Ihr Antivirusprogramm, um einen Komplettscan durchzuführen.
  • Translator 2: Bitte führen Sie mit Ihrem Virenschutzprogramm eine komplette Überprüfung durch.
• There are useful segments with fewer issues/errors.
• To date, we are not able to automatically identify translation errors.
Towards a Human-Informed HQMT Development Cycle

# Who needs MT-Evaluation?

<table>
<thead>
<tr>
<th>Means</th>
<th>Task-specific?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MT Researchers:</strong></td>
<td></td>
</tr>
<tr>
<td>• Rapid feedback for engineering.</td>
<td></td>
</tr>
<tr>
<td>• Which setting is better?</td>
<td></td>
</tr>
<tr>
<td>• Are differences significant?</td>
<td></td>
</tr>
<tr>
<td>Shallow surface comparison with one (!) reference translation</td>
<td></td>
</tr>
</tbody>
</table>

| Intrinsic |

| Extrinsic |
How humans can provide feedback

- Post-editing
- Analytic error annotation (MQM)
- Task-based evaluation
- Designing test suites
Automatic Post-Editing (APE)

- Experts post-edit MT output.
- Algorithms learn the post-edits.
- Example:

  **Source:** This option is available only for high (128-bit RC4 or AES) encryption.

  **Raw MT:** Diese Option ist nur verfügbar für hohe (128-Bit RC4 oder AES).

  **APE:** Diese Option ist nur verfügbar für hohe Verschlüsselung (128-Bit RC4 oder AES).

  **Reference:** Diese Option ist nur verfügbar für hohe Verschlüsselung (128-Bit RC4 oder AES).

(Example from Marco Turchi, FBK)
MQM annotation

- MQM = Multidimensional Quality Metrics
- Detailed error analysis
- Allows to create error profiles
- MQM/DQF standardisation initiative at ASTM
MQM annotation example

Go to Tools and then choose 'Delete browsing history..', you can then choose to delete your Internet cookies.

<table>
<thead>
<tr>
<th>DE_P1</th>
<th>Gehen Sie zu Tools und wählen Sie dann Browsingchronik Löschen..., können Sie dann vorziehen, Ihre Internet-Cookies zu löschen.</th>
</tr>
</thead>
</table>
|       | 1. Mistranslation [Gehen Sie zu]
|       | 2. Untranslated [Tools]
|       | 3. Mistranslation [Browsingchronik]
|       | 4. Part of speech [Löschen]
|       | 5. Word order [können]
|       | 6. Mistranslation [vorziehen]

<table>
<thead>
<tr>
<th>DE_P2</th>
<th>Sprung zu Extras und wählen Sie dann Browserverlauf Löschen..., Sie können dann Ihre Internet-Cookies Löschen.</th>
</tr>
</thead>
</table>
|       | 1. Mistranslation [Sprung zu]
|       | 2. Typography [..]
|       | 3. Omission [..]

CAT tools with plugins for the DQF Framework (thus DQF-MQM): Trados Studio, WorldServer, GlobalLink, SDLTMS, XTM, Kaleidoscope, translate5, and MateCat.

Arle Richard Lommel, Aljoscha Burchardt, Hans Uszkoreit

**Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics**

in: Attila Görög, Pilar Sánchez-Gijón (eds.): 3 Tradumàtica: tecnologies de la traducció volume 0 number 12, Pages 455-463, o.A., 12/2014

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Vivien Macketanz, Inguna Skadiņa, Matteo Negri, and Marco Turchi

**Translation Quality and Productivity: A Study on Rich Morphology Languages**

Machine Translation Summit XVI, Pages 55-71, Nagoya, Japan, Asia-Pacific Association for Machine Translation, 2017
## Error profiles by system and language

<table>
<thead>
<tr>
<th>Error type</th>
<th>DE–EN</th>
<th>EN–DE</th>
<th>EN–LV</th>
<th>EN–CS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PBMT</td>
<td>PBMT</td>
<td>NMT</td>
<td>PBMT</td>
</tr>
<tr>
<td>Accuracy</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>Addition</td>
<td>539</td>
<td>332</td>
<td>167</td>
<td>277</td>
</tr>
<tr>
<td>Mistranslation</td>
<td>487</td>
<td>967</td>
<td>852</td>
<td>274</td>
</tr>
<tr>
<td>Omission</td>
<td>516</td>
<td>690</td>
<td>355</td>
<td>395</td>
</tr>
<tr>
<td>Untranslated</td>
<td>278</td>
<td>162</td>
<td>24</td>
<td>79</td>
</tr>
<tr>
<td>Fluency</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>233</td>
</tr>
<tr>
<td>Grammar</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Function words</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Extraneous</td>
<td>302</td>
<td>525</td>
<td>245</td>
<td>49</td>
</tr>
<tr>
<td>Incorrect</td>
<td>139</td>
<td>804</td>
<td>449</td>
<td>56</td>
</tr>
<tr>
<td>Missing</td>
<td>362</td>
<td>779</td>
<td>231</td>
<td>66</td>
</tr>
<tr>
<td>Word form</td>
<td>0</td>
<td>94</td>
<td>267</td>
<td>280</td>
</tr>
<tr>
<td>Part of speech</td>
<td>20</td>
<td>128</td>
<td>132</td>
<td>38</td>
</tr>
<tr>
<td>Agreement</td>
<td>18</td>
<td>506</td>
<td>97</td>
<td>419</td>
</tr>
<tr>
<td>Tense/aspect/mood</td>
<td>63</td>
<td>184</td>
<td>51</td>
<td>60</td>
</tr>
<tr>
<td>Word order</td>
<td>218</td>
<td>868</td>
<td>309</td>
<td>336</td>
</tr>
<tr>
<td>Spelling</td>
<td>118</td>
<td>126</td>
<td>132</td>
<td>324</td>
</tr>
<tr>
<td>Typography</td>
<td>282</td>
<td>553</td>
<td>249</td>
<td>823</td>
</tr>
<tr>
<td>Unintelligible</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Terminology</td>
<td>27</td>
<td>82</td>
<td>139</td>
<td>34</td>
</tr>
<tr>
<td>All categories</td>
<td>3336</td>
<td>6775</td>
<td>3700</td>
<td>3803</td>
</tr>
</tbody>
</table>

Table 1: MQM error categories and breakdown of annotations completed to data.
TASK-BASED EVALUATION
Extrinsic Evaluation Scenario

End User

Retrieval

Publication

Question: Language X

Answer: Language X
Basis: The QTLeap Corpus

- **4000 interactions** (question-answer pairs), e.g.:
  - **Question-EN**: *What is the latest wireless standard?*
  - **Answer-EN**: *The latest standard is the norm N.*

- **8 languages** (X<->EN)
  - Basque, Bulgarian, Czech, Dutch, English, German, Portuguese and Spanish

- On META-SHARE
Evaluating the Publication step: Experiments

- **Pilot 0**: Estimating probability of calling operator
- **Pilot 1**: Comparison with Pilot 0
- **Pilot 2**: Ranking of three Pilots (WMT style)

- **Human evaluation**
  - At least three volunteers per language (no IT experts)
  - Evaluation on 100 interactions
  - Web forms
Step 1: Review answer A (MT) without any reference:
- It would clearly help me solve my problem / answer my question
- It might help, but would require some thinking to understand it.
- Is not helpful / I don't understand it

Step 2: Compare answers A and B (human reference), (re-)evaluate A selecting one of the following options:
- A gives the right advice.
- A gets minor points wrong.
- A gets important points wrong.
### Results of Step 1 and 2

<table>
<thead>
<tr>
<th>Statement</th>
<th>EU</th>
<th>BG</th>
<th>CS</th>
<th>NL</th>
<th>DE</th>
<th>PT</th>
<th>ES</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>It would clearly help me solve my problem / answer my question</td>
<td>30.7%</td>
<td>48.1%</td>
<td>49.5%</td>
<td>24.7%</td>
<td>37.3%</td>
<td>12.4%</td>
<td>65.3%</td>
<td>38.3%</td>
</tr>
<tr>
<td>It might help, but would require some thinking to understand it.</td>
<td>47.7%</td>
<td>43.6%</td>
<td>35.2%</td>
<td>43.4%</td>
<td>41.4%</td>
<td>35.3%</td>
<td>26.3%</td>
<td>39.0%</td>
</tr>
<tr>
<td>It is not helpful / I don’t understand it</td>
<td>21.7%</td>
<td>8.3%</td>
<td>15.3%</td>
<td>31.6%</td>
<td>21.3%</td>
<td>52.3%</td>
<td>8.3%</td>
<td>22.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statement</th>
<th>EU</th>
<th>BG</th>
<th>CS</th>
<th>NL</th>
<th>DE</th>
<th>PT</th>
<th>ES</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>A gives the right advice.</td>
<td>25.7%</td>
<td>35.0%</td>
<td>42.2%</td>
<td>25.6%</td>
<td>43.2%</td>
<td>22.9%</td>
<td>45.3%</td>
<td>34.3%</td>
</tr>
<tr>
<td>A gets minor points wrong.</td>
<td>37.7%</td>
<td>44.3%</td>
<td>31.9%</td>
<td>35.9%</td>
<td>33.4%</td>
<td>23.2%</td>
<td>22.3%</td>
<td>32.7%</td>
</tr>
<tr>
<td>A gets important points wrong.</td>
<td>36.7%</td>
<td>20.7%</td>
<td>25.9%</td>
<td>38.4%</td>
<td>23.4%</td>
<td>54.0%</td>
<td>32.3%</td>
<td>33.1%</td>
</tr>
</tbody>
</table>
## Estimating operator invention probability

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Step 2</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Solves my problem</td>
<td>Gets the right advice</td>
<td>low</td>
</tr>
<tr>
<td>B Solves my problem</td>
<td>Gets minor points wrong</td>
<td>low</td>
</tr>
<tr>
<td>C Would require some thinking to understand it</td>
<td>Gets the right advice</td>
<td>low</td>
</tr>
<tr>
<td>D Would require some thinking to understand it</td>
<td>Gets minor points wrong</td>
<td>medium</td>
</tr>
<tr>
<td>E Solves my problem</td>
<td>Gets important points wrong</td>
<td>high</td>
</tr>
<tr>
<td>F Would require some thinking to understand it</td>
<td>Gets important points wrong</td>
<td>high</td>
</tr>
<tr>
<td>G Is not helpful / I don't understand it</td>
<td>Gets the right advice</td>
<td>high</td>
</tr>
<tr>
<td>H Is not helpful / I don't understand it</td>
<td>Gets minor points wrong</td>
<td>high</td>
</tr>
<tr>
<td>I Is not helpful / I don't understand it</td>
<td>Gets important points wrong</td>
<td>high</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probability</th>
<th>EU</th>
<th>BG</th>
<th>CS</th>
<th>NL</th>
<th>DE</th>
<th>PT</th>
<th>ES</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>33.3%</td>
<td>47.4%</td>
<td>54.5%</td>
<td>30.4%</td>
<td>47.8%</td>
<td>21.5%</td>
<td>60.4%</td>
<td>42.2%</td>
</tr>
<tr>
<td>medium</td>
<td>28.1%</td>
<td>30.6%</td>
<td>17.9%</td>
<td>21.9%</td>
<td>22.0%</td>
<td>15.8%</td>
<td>7.0%</td>
<td>20.5%</td>
</tr>
<tr>
<td>high</td>
<td>37.0%</td>
<td>22.0%</td>
<td>27.5%</td>
<td>47.7%</td>
<td>30.1%</td>
<td>62.7%</td>
<td>32.7%</td>
<td>37.1%</td>
</tr>
</tbody>
</table>
Pilot 1: Direct comparison

- Supposed that the reference answer is correct, the evaluator is asked which of the two answers (A or B) provides a better answer to the question.
- The possible options are:
  - A is a better answer than B
  - B is a better answer than A
  - A and B are equally good answers
  - A and B are equally bad answers
## Pilot 1: Results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>EU</th>
<th>BG</th>
<th>CS</th>
<th>NL</th>
<th>DE</th>
<th>PT</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) P1 better than P0</td>
<td>10.0%</td>
<td>16.6%</td>
<td>40.5%</td>
<td>22.3%</td>
<td>17.3%</td>
<td>30.8%</td>
<td>1.8%</td>
</tr>
<tr>
<td>b) P0 better than P1</td>
<td>65.0%</td>
<td>22.2%</td>
<td>24.5%</td>
<td>39.3%</td>
<td>24.2%</td>
<td>13.0%</td>
<td>70.5%</td>
</tr>
<tr>
<td>c) P1 and P0 are equally good</td>
<td>2.5%</td>
<td>25.6%</td>
<td>20.0%</td>
<td>8.5%</td>
<td>22.1%</td>
<td>6.2%</td>
<td>2.6%</td>
</tr>
<tr>
<td>d) P1 and P0 are equally bad</td>
<td>22.5%</td>
<td>35.6%</td>
<td>15.0%</td>
<td>29.9%</td>
<td>36.4%</td>
<td>50.0%</td>
<td>25.1%</td>
</tr>
<tr>
<td>e) “P1 good”: a) + c)</td>
<td>12.5%</td>
<td>42.2%</td>
<td>60.5%</td>
<td>30.8%</td>
<td>39.4%</td>
<td>37.0%</td>
<td>4.4%</td>
</tr>
<tr>
<td>f) “P0 good”: b) + c)</td>
<td>67.5%</td>
<td>47.8%</td>
<td>44.5%</td>
<td>47.8%</td>
<td>46.3%</td>
<td>19.2%</td>
<td>73.1%</td>
</tr>
</tbody>
</table>
1. Die Frage, die Sie eben gelesen haben:

Die Bedeutung der Tastenkombination STRG + SHIFT + N (Google Chrome)?

2. Lesen Sie jetzt die Referenzantwort:

Es öffnet sich der Incognito-Modus. Es ermöglicht Ihnen, sich im Internet zu bewegen, ohne Informationen auf Ihrem PC zu speichern.

3. Lesen Sie diese drei alternativen Antworten und ordnen Sie sie von gut (1) nach schlecht (3).
Wenn Sie denken, dass zwei Antworten die gleiche Qualität haben, können Sie dieselbe Zahl mehrfach vergeben.
Zum Beispiel können Sie die Antworten A-B-C als 1-2-3 oder 2-1-3 oder 2-2-1 oder 1-1-1 oder jede andere Kombination dieser Zahlen bewerten, die Ihnen passend erscheinen.

A  Gut  1  2  3  Schlecht
Es öffnet den Inkognito-Modus. Es können Sie im Web surfen, ohne etwaige Informationen auf Ihrem Computer.

B  Gut  1  2  3  Schlecht
Es öffnet den Inkognitomodus. Es ermöglicht es Ihnen, sich das Web anzusehen, ohne Informationen auf Ihrem Rechner zu speichern.

C  Gut  1  2  3  Schlecht
Es öffnet den Inkognito-Modus. Es können Sie im Web surfen, ohne etwaige Informationen auf Ihrem Computer.
P2 vs. P0 (left) and P1 (right)
Correlation with intrinsic evaluation

Figure 9: Comparison of user evaluation results and BLEU scores for Pilot 2 and Pilot 0

TEST SUITES
How can we systematically reduce errors?

- Test suites are a familiar tool in NLP in areas such as grammar development.
- Idea: Use test suites in MT development.
- By test suite, we refer to a selected set of source-target pairs that reflects interesting or difficult cases (MWEs, long-distance, negation, terminology, etc.).
- In contrast to a “real-life” corpus with reference translations, the input in a test suite may well be made-up or edited to isolate and illustrate issues.
Using test suites

• Systematically evaluate and compare system(variant)s
  • Gets all 20 imparatives right
  • Gets half of the imparatives right
  • Gets no imparatives rights
  • …
• Guide system improvement / error reduction
• Testing can be local/partial
  • Lexical ambiguity (German “Gericht”; English “court” vs. “dish”)
  • Prefix verbs (English “picked up …”; German “hob … auf”)
• Build custom test suites for domain/task/job…
Non-verbal agreement
- Coreference
- Lexical ambiguity
- Gapping
- Pseudogapping
- External possessor
- Structural ambiguity
- Right node raising
- Internal possessor
- Sluicing
- Internal possessor
- Stripping
- VP-ellipsis

Ambiguity
- Focus particle
- Question tag
- Inversion
- Multiple connectors
- Negative inversion
- Piedpiping
- Polar question
- Preposition stranding
- Split infinitive
- Topicalization
- Wh-movement

Coordination & ellipsis
- Coreference
- Collocation
- Date
- Domainspecific term
- Location
- Measuring unit
- Proper name
- Nominal MWE
- Prepositional MWE
- Verbal MWE

False friends
- Focus particle
- Question tag
- Inversion
- Multiple connectors
- Negative inversion
- Piedpiping
- Polar question
- Preposition stranding
- Split infinitive
- Topicalization
- Wh-movement

Long distance dependency & interregressive
- Coreference
- Collocation
- Date
- Domainspecific term
- Location
- Measuring unit
- Proper name
- Nominal MWE
- Prepositional MWE
- Verbal MWE

MWE
- Focus particle
- Question tag
- Inversion
- Multiple connectors
- Negative inversion
- Piedpiping
- Polar question
- Preposition stranding
- Split infinitive
- Topicalization
- Wh-movement

Named entity & terminology
- Focus particle
- Question tag
- Inversion
- Multiple connectors
- Negative inversion
- Piedpiping
- Polar question
- Preposition stranding
- Split infinitive
- Topicalization
- Wh-movement

Negation
- Focus particle
- Question tag
- Inversion
- Multiple connectors
- Negative inversion
- Piedpiping
- Polar question
- Preposition stranding
- Split infinitive
- Topicalization
- Wh-movement

Punctuation
- Quotation marks
- Catenative verb
- Complex transitive
- Resultative
- Contact clause
- Gerund
- Passive voice

Special verb types
- Adverbial clause
- Conditional
- Case government
- Cleft sentence
- Ditransitive
- Middle voice
- Direct speech
- Imperative
- Indirect speech
- Intransitive
- Infinitive clause
- Modal
- Object clause
- Modal negated
- Pseudo-cleft clause
- Present progressive
- Relative clause
- Reflexive
- Subject clause
- Transitive

Verb valency
- Focus particle
- Question tag
- Inversion
- Multiple connectors
- Negative inversion
- Piedpiping
- Polar question
- Preposition stranding
- Split infinitive
- Topicalization
- Wh-movement

Verb tense/aspect/mood
- Focus particle
- Question tag
- Inversion
- Multiple connectors
- Negative inversion
- Piedpiping
- Polar question
- Preposition stranding
- Split infinitive
- Topicalization
- Wh-movement

Verb valency
- Focus particle
- Question tag
- Inversion
- Multiple connectors
- Negative inversion
- Piedpiping
- Polar question
- Preposition stranding
- Split infinitive
- Topicalization
- Wh-movement

14 Barrier Categories

~ 65 Barriers

Altogether ~5000 example sentences
### Exemplary test suite entries De-En

<table>
<thead>
<tr>
<th>Source</th>
<th>Category</th>
<th>Phenumenon</th>
<th>Target (raw)</th>
<th>Target (edited)</th>
<th>Positive token (indicative)</th>
<th>Negative token (indicative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena machte sich früh vom Acker.</td>
<td>MWE</td>
<td>Idiom</td>
<td>Lena [left the field early].</td>
<td>Lena left early.</td>
<td>left early</td>
<td>field</td>
</tr>
<tr>
<td>Lisa hat Lasagne gemacht, sie ist schon im Ofen.</td>
<td>Non-verbal agreement</td>
<td>Coreference</td>
<td>Lisa has made lasagna, [she] is already in the oven.</td>
<td>Lisa has made lasagna, it is already in the oven.</td>
<td>it</td>
<td>she</td>
</tr>
<tr>
<td>Ich habe der Frau das Buch gegeben.</td>
<td>Verb tense/aspect/mood</td>
<td>Dittransitive - perfect</td>
<td>I [have] the woman of the Book.</td>
<td>I have given the woman the book.</td>
<td>given the book to the woman, gave the book to the woman, given the woman the book, gave the book to the woman.</td>
<td></td>
</tr>
<tr>
<td>System</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-PBMT</td>
<td>Old (phrase-based) version of Google Translate (online, February 2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-NMT</td>
<td>New (neural) version of Google Translate (online, November 2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS-PBMT</td>
<td>Open-source phrase-based system (Moses) that uses a default configuration to serve as a baseline (only De-En)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFKI-NMT</td>
<td>Barebone neural system from DFKI, based on an encoder-decoder neural architecture with attention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ED-NMT</td>
<td>Neural system from U Edinburgh, system was built using the Nematus toolkit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RWTH-NMT</td>
<td>NMT-system from RWTH, makes use of subword units and has been finetuned to perform well on the IWSLT 2016 spoken language task (only De-En)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBMT</td>
<td>Commercial rule-based system Lucy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Test suite experiment – evaluation procedure

• So far: manual checking
• One phenomenon at a time, e.g.:
  – For ambiguity: Do I find the right sense, no matter what I find in the rest of the sentence?
  – For a prefix verb: Do I find both parts?
  – For an English question: Do I see the Wh-Word and two verbs?
  – For a verb paradigm “X has given Y to Z”: Is the sentence complete and correct?
  – …

• Count results

<table>
<thead>
<tr>
<th>Category</th>
<th>#</th>
<th>O-PBMT</th>
<th>O-NMT</th>
<th>RBMT</th>
<th>OS-PBMT</th>
<th>DFKI-NMT</th>
<th>RWTH-NMT</th>
<th>ED-NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguity</td>
<td>17</td>
<td>12%</td>
<td>35%</td>
<td>42%</td>
<td>24%</td>
<td>35%</td>
<td>12%</td>
<td>35%</td>
</tr>
<tr>
<td>Composition</td>
<td>11</td>
<td>27%</td>
<td>73%</td>
<td>55%</td>
<td>27%</td>
<td>45%</td>
<td>45%</td>
<td>73%</td>
</tr>
<tr>
<td>Coordination &amp; ellipsis</td>
<td>8</td>
<td>25%</td>
<td>100%</td>
<td>38%</td>
<td>25%</td>
<td>38%</td>
<td>63%</td>
<td>63%</td>
</tr>
<tr>
<td>False friends</td>
<td>5</td>
<td>40%</td>
<td>40%</td>
<td>20%</td>
<td>20%</td>
<td>20%</td>
<td>40%</td>
<td>20%</td>
</tr>
<tr>
<td>Function words</td>
<td>19</td>
<td>5%</td>
<td>68%</td>
<td>21%</td>
<td>11%</td>
<td>26%</td>
<td>68%</td>
<td>42%</td>
</tr>
<tr>
<td>LDD &amp; interrogative</td>
<td>66</td>
<td>12%</td>
<td>79%</td>
<td>62%</td>
<td>21%</td>
<td>36%</td>
<td>55%</td>
<td>52%</td>
</tr>
<tr>
<td>MWE</td>
<td>42</td>
<td>14%</td>
<td>36%</td>
<td>7%</td>
<td>21%</td>
<td>10%</td>
<td>12%</td>
<td>19%</td>
</tr>
<tr>
<td>NE &amp; terminology</td>
<td>25</td>
<td>48%</td>
<td>48%</td>
<td>40%</td>
<td>52%</td>
<td>40%</td>
<td>48%</td>
<td>40%</td>
</tr>
<tr>
<td>Negation</td>
<td>6</td>
<td>17%</td>
<td>83%</td>
<td>83%</td>
<td>17%</td>
<td>100%</td>
<td>67%</td>
<td>83%</td>
</tr>
<tr>
<td>Subordination</td>
<td>36</td>
<td>22%</td>
<td>58%</td>
<td>50%</td>
<td>31%</td>
<td>47%</td>
<td>42%</td>
<td>31%</td>
</tr>
<tr>
<td>Verb tense/aspect/mood</td>
<td>529</td>
<td>59%</td>
<td>80%</td>
<td>91%</td>
<td>52%</td>
<td>53%</td>
<td>74%</td>
<td>63%</td>
</tr>
<tr>
<td>Verb valency</td>
<td>32</td>
<td>16%</td>
<td>50%</td>
<td>44%</td>
<td>13%</td>
<td>47%</td>
<td>38%</td>
<td>50%</td>
</tr>
<tr>
<td>Sum</td>
<td>796</td>
<td>363</td>
<td>582</td>
<td>592</td>
<td>341</td>
<td>377</td>
<td>501</td>
<td>446</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>46%</td>
<td>73%</td>
<td>74%</td>
<td>43%</td>
<td>47%</td>
<td>63%</td>
<td>56%</td>
</tr>
</tbody>
</table>
(1) Source: Er hat einen Kater, weil er sehr tierlieb ist.
Reference: He has a cat because he is very fond of animals.

O-PBMT: He has a hangover, because he is very fond of animals.
O-NMT: He has a cat because he is very fond of animals.
RBMT: He has a tomcat because it is very animal-dear.
OS-PBMT: He has a hangover because it is an encounter.
DFKI-NMT: He has a kater because he is very animal.
RWTH-NMT: He has a hangover because he’s very animal.
ED-NMT: He has a hangover because he is very animal-loving.
<table>
<thead>
<tr>
<th>(2)</th>
<th>Source:</th>
<th>Warum hörte Herr Muschler mit dem Streichen auf?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference:</td>
<td>Why did Mr. Muschler stop painting?</td>
</tr>
<tr>
<td></td>
<td>O-PBMT:</td>
<td>Why heard Mr. Muschler on with the strike?</td>
</tr>
<tr>
<td></td>
<td>O-NMT:</td>
<td>Why did Mr. Muschler stop the strike?</td>
</tr>
<tr>
<td></td>
<td>RBMT:</td>
<td>Why did Mr. Muschler stop with the strike?</td>
</tr>
<tr>
<td></td>
<td>OS-PBMT:</td>
<td>Why was Mr Muschler by scrapping on?</td>
</tr>
<tr>
<td></td>
<td>DFKI-NMT:</td>
<td>Why did Mr. Muschler listen to the rich?</td>
</tr>
<tr>
<td></td>
<td>RWTH-NMT:</td>
<td>Why did Mr. Muschler listen to the stroke?</td>
</tr>
<tr>
<td></td>
<td>ED-NMT:</td>
<td>Why did Mr. Muschler stop with the stump?</td>
</tr>
</tbody>
</table>
Test suite experiment – examples: modal particle

(5)  Source: Kommst du denn?
Reference: Are you coming?

O-PBMT: You coming?
O-NMT: Are you coming?
RBMT: Do you come?
OS-PBMT: If you arrive?
DFKI-NMT: Do you not?
RWTH-NMT: Are you coming?
ED-NMT: Are you coming?
Test suite experiment – examples: wh-movement

(6) Source: Warum macht der Tourist drei Fotos?
Reference: Why does the tourist take three fotos?

<table>
<thead>
<tr>
<th>System</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>O-PBMT</td>
<td>Why does the tourist three fotos?</td>
</tr>
<tr>
<td>O-NMT</td>
<td>Why does the tourist make three fotos?</td>
</tr>
<tr>
<td>RBMT</td>
<td>Why does the tourist make three fotos?</td>
</tr>
<tr>
<td>OS-PBMT</td>
<td>Why does the tourist three fotos?</td>
</tr>
<tr>
<td>DFKI-NMT</td>
<td>Why does the tourist make three fotos?</td>
</tr>
<tr>
<td>RWTH-NMT</td>
<td>Why is the tourist taking three fotos?</td>
</tr>
<tr>
<td>ED-NMT</td>
<td>Why does the tourist make three fotos?</td>
</tr>
</tbody>
</table>
Test suite experiment – examples: MWE

(7) Source: Die Arbeiter müssten in den sauren Apfel beißen.
Reference: The workers would have to bite the bullet.

O-PBMT: The workers would have to bite the bullet.
O-NMT: The workers would have to bite into the acid apple.
RBMT: The workers would have to bite in the acid apple.
OS-PBMT: The workers would have to bite the bullet.
DFKI-NMT: Workers would have to bite in the acid apple.
RWTH-NMT: The workers would have to bite into the clean apple.
ED-NMT: The workers would have to bite in the acidic apple.
Source: Ich glaube, dass es auch nicht die amerikanische Position unterstützt.

Reference: I think that it does not support the American position either.

O-PBMT: […] it also does not support the US position.

O-NMT: […] it does not support the American position either.

RBMT: […] it does not support the American position either.

OS-PBMT: [...] it is also not the American position.

DFKI-NMT: […] it does not support the American position either.

RWTH-NMT: […] it does not support the American position.

ED-NMT: […] it does not support the American position either.
Test suite experiment – examples: relative clause

(10) Source: Wie kann ich die Farbe, mit der ich arbeite, ändern?
Reference: How can I change the color I am working with?

O-PBMT: How can I change the color with which I work to change?
O-NMT: How can I change the color with which I work?
RBMT: How can I change the color with which I work?
OS-PBMT: How can I change the colour, with whom I work, change?
DFKI-NMT: How can I change the color I work with?
RWTH-NMT: How can I change the color I work with?
ED-NMT: How can I change the color I work with?
Test suite experiment – examples: modal negated pluperfect subjunctive II

(11) Source:  Ich hätte nicht lesen gedurft.
Reference:  I would not have been allowed to read.

**O-PBMT:**  I would not have been allowed to read.
**O-NMT:**  I should not have read.
**RBMT:**  I would not have been allowed to read.
**OS-PBMT:**  I would not have read gedurft.
**DFKI-NMT:**  I would not have been able to read.
**RWTH-NMT:**  I wouldn’t have read.
**ED-NMT:**  I wouldn’t have read.
(12) Source: Der Manager besteht auf den Test.
Reference: The manager insists on the test.

O-PBMT: The manager is on the test.
O-NMT: The manager insists on the test.
RBMT: The manager insists on the test.
OS-PBMT: The manager is on the test.
DFKI-NMT: The manager is on the test.
RWTH-NMT: The manager is on the test.
ED-NMT: The manager is on the test.
Data preparation

Language Direction: DE->EN

Generate Sentences per Category/Phenomenon

Generate 5 sentences or % from each of these categories all

Generate 5 sentences or % from each phenomenon in these categories all

Generate 5 sentences or % from each of these phenomena all
## Evaluation

<table>
<thead>
<tr>
<th>ID data point</th>
<th>Source</th>
<th>Category</th>
<th>Phenomenon</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>00005001</td>
<td>Er hob die Tür aus den Angels.</td>
<td>Ambiguity</td>
<td>Lexical ambiguity</td>
<td>He lifted the door.</td>
</tr>
<tr>
<td>00005002</td>
<td>Die Tür hing schief in den Angels.</td>
<td>Ambiguity</td>
<td>Lexical ambiguity</td>
<td>The door hung obliquely in the hinges.</td>
</tr>
<tr>
<td>00005003</td>
<td>Die Angels der Tür quietschten beim Öffnen.</td>
<td>Ambiguity</td>
<td>Lexical ambiguity</td>
<td>The angling of the door creaked open.</td>
</tr>
</tbody>
</table>
Regular Expressions

Source: Sie fuhr das Auto ihres Mannes.

Translation: She drove her husband's car.

Positive Regex:

husband|spouse|hubb(y|ies)

Negative Regex:

(gentle)?m[ae]n|guy

Positive Tokens:

Negative Tokens:

Update rules and result

Discard changes
CUSTOM TEST SUITES
### Technical test suite example

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>PB-SMT</th>
<th>RBMT</th>
<th>RBMT improved</th>
<th>neural</th>
<th>sel. mech.</th>
</tr>
</thead>
<tbody>
<tr>
<td>imperatives</td>
<td>247</td>
<td>68%</td>
<td>79%</td>
<td>79%</td>
<td>74%</td>
<td>*73%</td>
</tr>
<tr>
<td>compounds</td>
<td>219</td>
<td>55%</td>
<td>87%</td>
<td><strong>85%</strong></td>
<td>51%</td>
<td>70%</td>
</tr>
<tr>
<td>“&gt;” separators</td>
<td>148</td>
<td><strong>99%</strong></td>
<td>39%</td>
<td>83%</td>
<td>93%</td>
<td>80%</td>
</tr>
<tr>
<td>quotation marks</td>
<td>431</td>
<td><strong>97%</strong></td>
<td>94%</td>
<td>75%</td>
<td>95%</td>
<td>80%</td>
</tr>
<tr>
<td>verbs</td>
<td>505</td>
<td>85%</td>
<td>93%</td>
<td><strong>93%</strong></td>
<td>90%</td>
<td>*90%</td>
</tr>
<tr>
<td>phrasal verbs</td>
<td>90</td>
<td>22%</td>
<td>68%</td>
<td><strong>77%</strong></td>
<td>38%</td>
<td>53%</td>
</tr>
<tr>
<td>terminology</td>
<td>465</td>
<td><strong>64%</strong></td>
<td>50%</td>
<td>53%</td>
<td>55%</td>
<td>54%</td>
</tr>
<tr>
<td>sum</td>
<td>2105</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td></td>
<td>76%</td>
<td>77%</td>
<td>77%</td>
<td>75%</td>
<td>74%</td>
</tr>
</tbody>
</table>

Recent study on customer data

• Adopted Moses vs. unadopted NMT

<table>
<thead>
<tr>
<th>#</th>
<th>NMT</th>
<th>Moses</th>
</tr>
</thead>
<tbody>
<tr>
<td>formal address</td>
<td>138</td>
<td>90%</td>
</tr>
<tr>
<td>genitive</td>
<td>114</td>
<td>92%</td>
</tr>
<tr>
<td>modal construction</td>
<td>290</td>
<td>94%</td>
</tr>
<tr>
<td>negation</td>
<td>101</td>
<td>93%</td>
</tr>
<tr>
<td>passive voice</td>
<td>109</td>
<td>83%</td>
</tr>
<tr>
<td>predicate adjective</td>
<td>122</td>
<td>81%</td>
</tr>
<tr>
<td>prepositional phrase</td>
<td>104</td>
<td>81%</td>
</tr>
<tr>
<td>terminology</td>
<td>330</td>
<td>35%</td>
</tr>
<tr>
<td>tagging</td>
<td>145</td>
<td>83%</td>
</tr>
</tbody>
</table>

sum       1453

average   89%   73%

Table 2: Manual evaluation translation accuracy focusing on particular phenomena.

Conclusions

- Current evaluation workflow based on reference translation (and scores like BLEU) provides little insights about MT quality and the nature of errors
- Alternatives are being actively researched:
  - Learning from post-edits
  - Target analytics: Error annotation with MQM
  - Task-based evaluation
  - Source-driven testing: Test suites
  - Quality estimation, better automatic metrics, etc.
- Still: communication between communities (MT development and language experts) can be intensified
Thank you