Biases and perils of MT evaluation

Martin Popel
ÚFAL (Institute of Formal and Applied Linguistics)
Charles University, Prague

November 19th 2019, Workshop on document level MT evaluation, Luxembourg
In my presentation, I focus on the following aspects of evaluations: sentence-level vs. document-level vs. document-aware, source-based vs. reference-based, direct assessment vs. comparison-based, fluency-biased vs. adequacy-biased. I also discuss the aspect of translationese and native target/source-language translators and evaluators. Finally, I will present some results from the WMT19 Metrics shared task.
Perils of human translation

- Fluency vs. Adequacy

Fluency vs. Adequacy in translation is not always achievable.
Perils of human translation

- Fluency vs. Adequacy
- Intent ⇔ source language ⇔ translation

Fluency: not always achievable in source in translation

Adequacy
Perils of human translation

• Fluency vs. Adequacy
• Intent $\Rightarrow$ source language $\Rightarrow$ translation

• Semantics vs. Pragmatics. Example:
  • SRC: I'm not going to worry too much about it.
  • REF: I believe everything will be OK.
Perils of human translation

- Fluency vs. Adequacy
- Intent $\Rightarrow$ source language $\Rightarrow$ translation

- Semantics vs. Pragmatics. Example:
  - SRC: I'm not going to worry too much about it.
  - REF: I believe everything will be OK.

<table>
<thead>
<tr>
<th>not worry</th>
<th>worry</th>
</tr>
</thead>
<tbody>
<tr>
<td>believe OK</td>
<td>usual</td>
</tr>
<tr>
<td>don't believe OK</td>
<td>rare</td>
</tr>
</tbody>
</table>
Types of manual MT evaluation

- REF-based … show candidate and (human) reference
- SRC-based … show candidate and source sentence
- REF&SRC-based … show both
Types of manual MT evaluation

• REF-based … show candidate and (human) reference
• SRC-based … show candidate and source sentence
• REF&SRC-based … show both

• RR = Relative Ranking … relative, ordinal, N systems
• DA = Direct Assessment … “absolute”, continuous, 1 system
• RankME = rank-based magnitude estimation … continuous, N sys
Types of manual MT evaluation

- REF-based … show candidate and (human) reference
- SRC-based … show candidate and source sentence
- REF&SRC-based … show both

- Sentence-level
- Document-level … single score per document
- Document-aware … show whole documents, scores per sentence

- RR = Relative Ranking … relative, ordinal, N systems
- DA = Direct Assessment … “absolute”, continuous, 1 system
- RankME = rank-based magnitude estimation … continuous, N sys
Example: REF&SRC sent-level RR (WMT10–16)

"Valentino měl vždycky raději eleganci než slávu.
— Source

Valentino has always preferred elegance to notoriety.
— Reference

Best ← Rank 1 ← Rank 2 ← Rank 3 ← Rank 4 ← Rank 5 → Worst

"Valentino should always elegance rather than fame.
— Translation 1

Best ← Rank 1 ← Rank 2 ← Rank 3 ← Rank 4 ← Rank 5 → Worst

"Valentino has always rather than the elegance of glory.
— Translation 2

Best ← Rank 1 ← Rank 2 ← Rank 3 ← Rank 4 ← Rank 5 → Worst

" Valentino had always preferred elegance than glory.
— Translation 3

Best ← Rank 1 ← Rank 2 ← Rank 3 ← Rank 4 ← Rank 5 → Worst

"Valentino has always had the elegance rather than glory.
— Translation 4

Best ← Rank 1 ← Rank 2 ← Rank 3 ← Rank 4 ← Rank 5 → Worst

" Valentino has always had a rather than the elegance of the glory.
— Translation 5
Example: SRC doc-level DA (WMT19)
Example: pseudo doc-aware DA (WMT19)

For the pair of sentences below: Read the text and state how much you agree that:

The black text adequately expresses the meaning of the gray text in German (deutsch).

North Korea says 'no way' will disarm unilaterally without trust
— Source text

Nordkorea sagt, Sprünge ohne Vertrauen entwaffnen ohne Vertrauen.
— Candidate translation

0%  |  100%
Example: **pseudo doc-aware DA (WMT19)**

<table>
<thead>
<tr>
<th>Sentence pair</th>
<th>WMT19DocSrcDA #281:Document #reuters.218861-0</th>
<th>English → German (deutsch)</th>
</tr>
</thead>
</table>

For the pair of **sentences** below: Read the text and state how much you agree that:

> The black text adequately expresses the meaning of the gray text in German (deutsch).

**Source text**

North Korea says 'no way' will disarm unilaterally without trust

— Source text

**Candidate translation**

Nordkorea sagt, Sprünge ohne Vertrauen entwaffnen ohne Vertrauen.

— Candidate translation

0% | 100%

---

sentences in doc order, **but one sentence per screen and no undo/back button**
Example: SRC doc-aware 10-RankME

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Source</td>
<td>Translation1</td>
<td>TL_overall</td>
<td>TL_instructions</td>
<td>TL_quality</td>
<td>Translation2</td>
<td>TL_overall</td>
<td>TL_instructions</td>
<td>TL_quality</td>
<td>Optional comment</td>
</tr>
<tr>
<td>165</td>
<td>&quot;And we’re protecting our shareholders from employment litigation.&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>190</td>
<td>Companies started taking ethics, values and employee engagement more seriously in 2002 after accounting firm Arthur Andersen collapsed because of ethical violations from the Enron scandal, Quinlan said.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>170</td>
<td>But it wasn’t until “social media came into its own” that companies realized they couldn’t stop their dirty laundry from going viral online.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>171</td>
<td>&quot;Prior to using technology to monitor ethics, people used hope as a strategy,&quot; he said.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>172</td>
<td>Both Glint and Convercent offer their software as a service, charging companies recurring fees to use their products.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>173</td>
<td>It’s a business model and opportunity that has the approval of venture capital investors, who have propped up both start-ups.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>174</td>
<td>Convercent raised $10 million in funding in February from firms such as Sapphire Ventures and Tola Capital, bringing its total capital raised to $47 million.</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T1: chybný překlad termínů &quot;venture capital&quot;</td>
</tr>
<tr>
<td>175</td>
<td>Glint secured $10 million in November from Bessemer Venture Partners, bringing its total funding to $56 million.</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>These investments hardly come as a surprise, given the interconnected nature of companies, culture and venture capital.</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>177</td>
<td>There’s a growing body of research showing today’s employees expect more from their workplaces than before.</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>178</td>
<td>In competitive markets such as Silicon Valley, high salaries and interesting projects are merely table stakes.</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>179</td>
<td>Employees want to feel that they’re accepted and valued and that they’re giving true time to a company with a positive mission.</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Zaměstnanci chcějí mít pocit, že jsou přijímáni a ceniční a že věnují svůj čas společnosti, která ušlechtila v pozytivním posílání.

Podívejte se na kompetenční třídu, jakou je Silicon Valley, je vysoká přijímací řízení vysoké platy a zajímavé projekty.

Zaměstnanci chcějí mít pocit, že jsou přijímáni a ceniční a že věnují svůj čas společnosti s pozytivním posíláním.
Perils of manual MT evaluation

Each type of evaluation is biased towards some systems.

- REF-based … similarity to human errors (or post-editing)
- SRC-based … problems with non-professional evaluators
- REF&SRC-based … both

- Sent-level … false positives and false negatives (fluency+adeq.)
- Doc-level … too coarse, psychological problems
- Doc-aware … how to approximate doc-level? Avg, min…?

- RR … tiny improvements/errors same as big ones
- DA … fluency and serious adequacy errors only (but faster)
- RankME … slower, difficult if N>3
Which system is better? Median vs Mean?

- Reference: mean: 8.3, median: 8.5
- CUBBITT: mean: 8.2, median: 9.0

$P = 0.587$

<table>
<thead>
<tr>
<th>Reference better</th>
<th>Tie</th>
<th>Transformer better</th>
</tr>
</thead>
<tbody>
<tr>
<td>38%</td>
<td>21%</td>
<td>41%</td>
</tr>
</tbody>
</table>
Perils of automatic evaluation

BLEU (& other REF-similarity metrics) has 3 issues:
• Not enough REFs, i.e. low coverage of correct translations.
• Differences in BLEU do not correlate with Human scores (even with enough REFs).
• Human REFs may be worse than MT.

If REF is a translation:
• It may not be adequate and fluent (if non-professional translator).

If REF is the original sentence (reverse-direction eval):
• SRC is not original, thus may not have the same meaning as REF
• SRC is likely not representative of the expected use case (domain/country & translationese)
• (tiny) risk of non-perfect adequacy+fluency of REF relative to the intent
BLEU does not correlate with humans for strong systems

From WMT19 Metrics task (Ma et al., 2019), EN-DE
Domain&orig-lang effect: BLEU WMT18 EN→DE

sacrebleu --detail
Is NMT better in fluency or adequacy?
Is NMT better in fluency or adequacy?
Is NMT better in fluency or adequacy?

- **Adequacy (n = 346)**
  - Mean: 8.1
  - Median: 8.0
  - Reference: 26%
  - CUBBITT: 52%
  - **P = 4.6e-08 (***)**

- **Fluency (n = 346)**
  - Mean: 9.0
  - Median: 9.5
  - Reference: 48%
  - CUBBITT: 26%
  - **P = 2.1e-06 (***)**
Domain effect: manual doc-level adequacy

<table>
<thead>
<tr>
<th>Category</th>
<th>n</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>business</td>
<td>5</td>
<td>0.625</td>
</tr>
<tr>
<td>crime</td>
<td>9</td>
<td>1.000</td>
</tr>
<tr>
<td>entertainment</td>
<td>5</td>
<td>0.625</td>
</tr>
<tr>
<td>politics</td>
<td>12</td>
<td>0.109</td>
</tr>
<tr>
<td>scitech</td>
<td>7</td>
<td>1.000</td>
</tr>
<tr>
<td>sport</td>
<td>9</td>
<td>0.727</td>
</tr>
<tr>
<td>world</td>
<td>6</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Domain effect: manual sent-level adequacy
Domain effect: manual doc-level fluency

<table>
<thead>
<tr>
<th>Category</th>
<th>n</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>business</td>
<td>5</td>
<td>1.000</td>
</tr>
<tr>
<td>crime</td>
<td>9</td>
<td>0.016 *</td>
</tr>
<tr>
<td>entertainment</td>
<td>5</td>
<td>0.375</td>
</tr>
<tr>
<td>politics</td>
<td>12</td>
<td>0.227</td>
</tr>
<tr>
<td>scitech</td>
<td>7</td>
<td>0.219</td>
</tr>
<tr>
<td>sport</td>
<td>9</td>
<td>0.004 **</td>
</tr>
<tr>
<td>world</td>
<td>6</td>
<td>0.063</td>
</tr>
</tbody>
</table>
Domain effect: manual sent-level fluency

<table>
<thead>
<tr>
<th>Category</th>
<th>n</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>business</td>
<td>45</td>
<td>0.441</td>
</tr>
<tr>
<td>crime</td>
<td>90</td>
<td>3.8e-04</td>
</tr>
<tr>
<td>entertainment</td>
<td>49</td>
<td>5.7e-08</td>
</tr>
<tr>
<td>politics</td>
<td>115</td>
<td>4.7e-04</td>
</tr>
<tr>
<td>scitech</td>
<td>65</td>
<td>0.063</td>
</tr>
<tr>
<td>sport</td>
<td>90</td>
<td>7.9e-11</td>
</tr>
<tr>
<td>world</td>
<td>58</td>
<td>0.003</td>
</tr>
</tbody>
</table>

CUBBITT - Reference

Sentence-level fluency

- Median
- Mean
Overall quality = $x \cdot \text{adequacy} + (1-x) \cdot \text{fluency}$?
Overall quality = x*adequacy + (1-x)*fluency?
Overall quality = x*adequacy + (1-x)*fluency?
Perils of adequacy w.r.t. purpose/localization

_Burning Man → Matějská pout’ (St. Matthew‘s Funfair)_