A journey of MT research

Why it is crucial to understand evaluation

Markus Freitag
MT Marathon 2022
**Model**

APE at Scale

*Goal*: Improve naturalness of the MT output

*Outcome*: Automatic Metrics are biased towards literal, non-natural output

---

**Automatic Eval**

BLEU might be Guilty/ WMT20 Metric Task

*Goal*: Unbiased automatic metrics

*Outcome*: Human raters prefer easy-to-explain output

---

**Human Eval**

Expert-based Human Eval

*Goal*: Reliable, explainable human evaluation

*Outcome*: Reliable, explainable human evaluation

---

**Automatic Eval**

WMT21 Metric Task

*Goal*: Re-evaluate with better ground truth

*Outcome*: Neural metrics correlate well with updated evaluation protocol

---

**Model**

Natural Diet/ MBR-Decoding

*Goal*: Find interesting novel translation approaches

*Outcome*: Improvements only visible with updated evaluation protocol

---

2019

**Briefly**

(Outline)

2022

**Briefly**

(Outline)

**In Detail**

(Outline)
Should we question the metric?

#1

How we start questioning BLEU
APE at Scale and its Implications on MT Evaluation Biases [WMT19]

**Goal**: Improve naturalness of the MT output

**2nd goal**: Improve accuracy (the APE sees the full translation)

**Approach**: Automatic post-editing on synthetic data

**Outcome**

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT</td>
<td>34.3</td>
<td>4.64</td>
</tr>
<tr>
<td>MT+APE</td>
<td>30.7</td>
<td>4.63</td>
</tr>
</tbody>
</table>

**Automatic Metrics**: negative

**Crowd Scalar-Value Human Eval**: neutral

**But - Impression**: MT+APE generates more natural, and accurate translations. Problems with evaluation?
What’s wrong with BLEU?
A quick detour
What BLEU actually measures?
BLEU might be Guilty but References are not Innocent

Top matching 4-grams of Facebook with WMT reference:

1. , sagte er → 28 times (, he said.)
2. “, sagte er → 14 times (” , he said)
3. fuegte hinzu , dass → 8 times (added that)

→ Easy way to generate translation output with high BLEU score:

1. Match the ngrams responsible for the sentence structure [translating literally gives you the highest success rate - never ever be creative or change the structure!]
2. Translate as simple as possible. Using frequent words have a high chance to find a counterpart in the reference translation.

→ BT, LLMs, MBR Decoding, APE models are typical approaches that improve the output by being more creative and thus yield low BLEU scores.
Should we question the metric?

#2

How we start questioning human evaluation
**Translationese as a Language in “Multilingual” NMT + WMT20 Metric Task**

**Goal**: Steer Model towards training data with more natural target

**Approach**: Classify Training data and focus on training examples with natural target

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>Human Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT</td>
<td>44.6</td>
<td>4.67</td>
</tr>
<tr>
<td>NAT MT</td>
<td>41.5</td>
<td>4.72</td>
</tr>
</tbody>
</table>

Similar observation as before. Broken Eval?

**WMT20 Metric Task**

<table>
<thead>
<tr>
<th></th>
<th>Rank of human translations</th>
<th>Kendall Tau Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-De</td>
<td>1, 4, 10</td>
<td></td>
</tr>
<tr>
<td>Zh-En</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

**Zh->En**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Kendall Tau Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMET</td>
<td>0.28</td>
</tr>
<tr>
<td>BLEURT</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Natural translations, in particular human translations are heavily penalized!
Human Evaluation
Why Human Evaluation?

- Human Eval
- Automatic Metrics
- Test Sets
- MT
- Training Data

Everything depends on the quality of human evaluation!
How to Measure the Quality of Human Evaluations

There is no ground truth!
Evaluation Based on Anchor Points

Human Eval

Q(Human Translation) ≥ Q(Machine Translation)

Q(Domain-adapted MT) ≥ Q(Open Domain MT)

Current practices violate these anchor points for high quality language pairs.
Additional Motivation beyond Reliability of System-level scores

**Feedback**

- Current practices return a scalar value per segment/system
  - How to interpret the score?
  - Is an improvement real or just rating noise?
- Better feedback:
  - Give details about error categories and severities
  - Is your method really doing what you intended?
  - Helps guide research!
  - Choose error weights for your task

**Automatic Metrics**

- Automatic metrics heavily rely on segment-level ratings
  - For training and evaluation
- Currently they are noisy and the general impression is that they are very noisy
  - No conclusive answer when comparing 2 metrics
  - Blocker for research in automatic metrics
Current Practices: Example WMT

- Segment-level ratings with document context (SR+DC) on a 0-100 scale
- Out-of-English:
  - Source-based
  - Rater pool: researchers/translators
- Into-English:
  - Reference-based
  - Rater pool: crowd-workers
- Rater quality control
  - Remove bad ratings
  - Not all segments get a rating
- Z-normalize ratings
  - Put raters on the same scale
Okay, convinced? What should we do?
Goals of this study:

1. Adapt standards from the (human) translator world
2. Re-evaluate current popular approaches
3. Give recommendations on how to conduct reliable human evaluation
4. Re-evaluate quality of automatic metrics
5. Define current error types in machine translation output
Multidimensional Quality Metrics (MQM)
Multidimensional Quality Metrics (MQM)

- Developed in the EU QTLaunchPad and QT21 projects (www.qt21.eu)
- Generic framework for assessing translation quality, adaptable to various evaluation needs - standard error hierarchy
- Fairly widely adopted by LSPs to evaluate MT and HT. Not so widely adopted by MT researchers.
- To use:
  - Choose relevant errors
  - Choose severity levels
  - Specify weights on errors and severities

MQM

- Accuracy: addition, mistranslation, omission, untranslated
- Fluency: grammar, inconsistency, spelling, typography, unintelligible
- Design: style, terminology, completeness
- Locale: convention, locale-specific content

MQM

To use:
- Choose relevant errors
- Choose severity levels
- Specify weights on errors and severities
MQM Demo
Adapting MQM for Broad-Coverage MT (Translate)

Our MQM schema

- Standard top-level errors: **Accuracy**, **Fluency**, **Terminology**, **Style**, and **Locale** - dedicated sub-categories
- Special error category for completely garbled output: **Non-translation!**
- Three severities:
  - **Major**: real errors
  - **Minor**: imperfections
  - **Neutral**: rater vent
- All standard errors have equal weight except easily-fixable presentation errors

Error Weighting

- Maximum Error count per segment: 25
- System-level error count is the sum of all errors

<table>
<thead>
<tr>
<th>Severity</th>
<th>Category</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major</td>
<td>Non-translation all other</td>
<td>25</td>
</tr>
<tr>
<td>Minor</td>
<td>Fluency/Punctuation all other</td>
<td>0.1</td>
</tr>
<tr>
<td>Neutral</td>
<td>all</td>
<td>0</td>
</tr>
</tbody>
</table>
Why is Error Annotation Superior to Asking for a Scalar Value?

Main idea:
- When annotators assign scores or rank translations, their decisions are (or should be?) implicitly based on identifying errors and other imperfections.
- Grounding assessments in explicit error identification creates a “platinum standard” for human evaluation.
  - New, simpler/cheaper schemas can measure correlations to this platinum dataset

Advantages:
- No temptation to “wing it” on long or complex segments
  - Annotators have to “explain” their ratings
  - Fair evaluation of more creative translations!
- Access to annotator rationale, for standardizing ratings and improving systems
- Weight different errors differently depending on the task not on the rater
  - “Taking away the burden of scoring errors”
  - Errors can be differently important for different tasks
Rater Pool - Crowd vs. Prof. Translator

Crowd/Researcher (WMT)

- Pro:
  - Large rater pool
  - Fast evaluation
  - Impact of one rater is minimal
- Cons:
  - Segment-level ratings noisy
  - Needs rater quality control
  - Biases:
    - Prefer the easy, direct translations

Professional Translator (MQM)

- Pro:
  - Native in the target language
  - Fluent in the source language
  - Are trained for the task
  - Reliable segment-level ratings
- Cons:
  - Small rater pool
    - One rater can have a large impact on the final result
  - More expensive
  - Slower turnaround time
Experiments
Experimental Setup - WMT 2020

WMT Submissions
- Top submission of WMT2020
- 6 for ZhEn, 5 for EnDe
- Very similar systems
- Domain-adapted systems
- The best translations we can generate for the news-domain
- Used for research

Online Systems
- 2 online systems:
  - Online-A
  - Online-B

Human Translations
- 2 standard human translations (Human-A, Human-B)
- 1 Paraphrased translation for EnDe (Human-P)
- Generated in-context

WHY
- Different training data
- Not tuned on news-translations
- Worse quality on news domain -> Should be ranked last

WHY
- The future of MT
- Should be ranked ahead of MT
### English->German System-level Rankings

<table>
<thead>
<tr>
<th>System</th>
<th>WMT↑</th>
<th>MQM↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-B</td>
<td>0.57(1)</td>
<td>0.75(1)</td>
</tr>
<tr>
<td>Human-A</td>
<td>0.45(4)</td>
<td>0.91(2)</td>
</tr>
<tr>
<td>Human-P</td>
<td>0.30(10)</td>
<td>1.41(3)</td>
</tr>
<tr>
<td>Tohoku-AIP-NTT</td>
<td>0.47(3)</td>
<td>2.02(4)</td>
</tr>
<tr>
<td>OPPO</td>
<td>0.50(2)</td>
<td>2.25(5)</td>
</tr>
<tr>
<td>eTranslation</td>
<td>0.31(9)</td>
<td>2.33(6)</td>
</tr>
<tr>
<td>Tencent_Translation</td>
<td>0.39(6)</td>
<td>2.35(7)</td>
</tr>
<tr>
<td>Huoshan_Translate</td>
<td>0.33(7)</td>
<td>2.45(8)</td>
</tr>
<tr>
<td>Online-B</td>
<td>0.42(5)</td>
<td>2.48(9)</td>
</tr>
<tr>
<td>Online-A</td>
<td>0.32(8)</td>
<td>2.99(10)</td>
</tr>
</tbody>
</table>

- MQM correctly ranks the anchor points
  - Human translations are ranked higher than MT
  - Human-P are paraphrased translations - very challenging to evaluate
- WMT has low correlation with MQM
  - Revising the system ranking in WMT20
Impact on Automatic Evaluation: Metrics from WMT20 Task

- Correlation of metrics is very different when comparing to MQM (orange) vs WMT (blue).
- Dotted line is MQM/WMT (human/human) correlation.
- Most metrics outperform WMT human scores!

<table>
<thead>
<tr>
<th>Zh-&gt;En</th>
<th>WMT ratings</th>
<th>MQM ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMET</td>
<td>0.28</td>
<td>0.71</td>
</tr>
<tr>
<td>BLEURT</td>
<td>0.07</td>
<td>0.64</td>
</tr>
<tr>
<td>WMT</td>
<td>n/a</td>
<td>0.28</td>
</tr>
</tbody>
</table>
### Error Category Distribution

<table>
<thead>
<tr>
<th>Error Categories</th>
<th>Human MQM</th>
<th>All MT MQM vs H.</th>
<th>Tohoku MQM vs H.</th>
<th>OPPO MQM vs H.</th>
<th>eTrans MQM vs H.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy/Misttransl</strong></td>
<td>0.296</td>
<td>1.285 4.3</td>
<td>1.026 3.5</td>
<td>1.219 4.1</td>
<td>1.244 4.2</td>
</tr>
<tr>
<td>Style/Awkward</td>
<td>0.146</td>
<td>0.299 2.0</td>
<td>0.289 2.0</td>
<td>0.315 2.1</td>
<td>0.296 2.0</td>
</tr>
<tr>
<td>Fluency/Grammar</td>
<td>0.097</td>
<td>0.224 2.3</td>
<td>0.193 2.0</td>
<td>0.215 2.2</td>
<td>0.196 2.0</td>
</tr>
<tr>
<td><strong>Accuracy/Omission</strong></td>
<td>0.070</td>
<td>0.091 1.3</td>
<td>0.063 0.9</td>
<td>0.063 0.9</td>
<td><strong>0.120 1.7</strong></td>
</tr>
<tr>
<td>Accuracy/Addition</td>
<td>0.067</td>
<td>0.025 0.4</td>
<td>0.018 0.3</td>
<td>0.024 0.4</td>
<td>0.021 0.3</td>
</tr>
<tr>
<td>Terminology/Inappropriate</td>
<td>0.061</td>
<td>0.193 3.2</td>
<td>0.171 2.8</td>
<td>0.189 3.1</td>
<td>0.193 3.2</td>
</tr>
<tr>
<td>Fluency/Spelling</td>
<td>0.050</td>
<td>0.039 1.3</td>
<td>0.030 1.0</td>
<td>0.039 1.3</td>
<td>0.028 0.9</td>
</tr>
<tr>
<td>Accuracy/Untranslated text</td>
<td>0.024</td>
<td>0.090 3.8</td>
<td>0.082 3.5</td>
<td>0.066 2.8</td>
<td>0.098 4.2</td>
</tr>
<tr>
<td>Fluency/Punctuation</td>
<td>0.014</td>
<td>0.039 2.8</td>
<td>0.067 4.9</td>
<td>0.013 1.0</td>
<td>0.011 0.8</td>
</tr>
<tr>
<td>Other</td>
<td>0.005</td>
<td>0.010 1.9</td>
<td>0.009 1.6</td>
<td>0.010 1.9</td>
<td>0.007 1.2</td>
</tr>
<tr>
<td>Fluency/Register</td>
<td>0.005</td>
<td>0.014 3.0</td>
<td>0.009 1.9</td>
<td>0.015 3.2</td>
<td>0.015 3.3</td>
</tr>
<tr>
<td>Terminology/Inconsistent</td>
<td>0.004</td>
<td>0.005 1.2</td>
<td>0.004 0.9</td>
<td>0.005 1.2</td>
<td>0.005 1.2</td>
</tr>
<tr>
<td>Non-translation</td>
<td>0.003</td>
<td>0.083 28.5</td>
<td>0.041 14.6</td>
<td>0.065 22.0</td>
<td><strong>0.094 32.0</strong></td>
</tr>
<tr>
<td>Fluency/Inconsistency</td>
<td>0.003</td>
<td>0.002 0.7</td>
<td>0.001 0.3</td>
<td>0.001 0.3</td>
<td>0.003 1.0</td>
</tr>
<tr>
<td>Fluency/Character enc.</td>
<td>0.002</td>
<td>0.001 0.7</td>
<td>0.002 1.0</td>
<td>0.001 0.6</td>
<td>0.000 0.2</td>
</tr>
<tr>
<td>All accuracy</td>
<td>0.457</td>
<td>1.492 3.3</td>
<td>1.189 2.6</td>
<td>1.372 3.0</td>
<td>1.483 3.2</td>
</tr>
<tr>
<td>All fluency</td>
<td>0.150</td>
<td>0.320 2.1</td>
<td>0.303 2.0</td>
<td>0.284 1.9</td>
<td>0.253 1.7</td>
</tr>
<tr>
<td>All except acc. &amp; fluenc.</td>
<td>0.222</td>
<td>0.596 2.7</td>
<td>0.526 2.4</td>
<td>0.591 2.7</td>
<td>0.596 2.7</td>
</tr>
<tr>
<td>All categories</td>
<td>0.829</td>
<td>2.408 2.9</td>
<td>2.017 2.4</td>
<td>2.247 2.7</td>
<td>2.332 2.8</td>
</tr>
</tbody>
</table>

MQM gives feedback!

1. **Tohoku:**
   - Fewer mistranslations
   - More punctuation errors

2. **eTrans:**
   - More Omission errors
   - More Non-Translations!
Impact of Improved Human Evaluation Protocol
WMT21 Metric Task

Changes:

- Jointly with Unbabel:
  - Expert-based human ratings of WMT submissions with MQM for 3 language pairs
- Addition of interesting metric systems that are challenging for both humans and machines
- Evaluation beyond the mean metric scores!

Goal:

- Establish standard + tooling for both human and automatic eval

WMT21 Results

Results in line with the ones of WMT20

- MQM ranks human translations higher than MT and correlates much better with metrics
- WMT DA scores correlate poorly with MQM and metrics

Use MQM annotation for metric research

- We would like to encourage everyone working on metrics to use the MQM annotation as groundtruth.
- All data is available on github!
A Natural Diet: Towards Improving Naturalness of Machine Translation Output [ACL 22]
An Example of Error-driven Research

**Experimental Human Results**

<table>
<thead>
<tr>
<th>MQM EnDe (extract)</th>
<th>base</th>
<th>&lt;nat&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major</td>
<td>44</td>
<td>51</td>
</tr>
<tr>
<td>Minor</td>
<td>79</td>
<td>26</td>
</tr>
<tr>
<td>Major</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Minor</td>
<td>56</td>
<td>29</td>
</tr>
<tr>
<td>Awkward</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>Total Errors</td>
<td>113</td>
<td>132</td>
</tr>
<tr>
<td>Total Errors</td>
<td>395</td>
<td>275</td>
</tr>
<tr>
<td>Global Score</td>
<td>0.91</td>
<td>0.88</td>
</tr>
</tbody>
</table>

**Approach**

- **Natural target**
  - Add `<nat>` tag

- **Translation target**
  - Add `<trans>` tag

**Inference:** Add `<nat>` tag to the source sentence

- Shows that method works!
- But: Introduces new errors (see major-mistranslation)

**Awkward category significantly improved, while mean score is almost neutral**
The impact of better understanding
Minimum Bayes Risk Decoding with Neural Metrics
High Quality Rather than High Model Probability
## Motivation: MAP Decode for Translation Task

<table>
<thead>
<tr>
<th>system</th>
<th>translations</th>
<th>log pplx</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Der Ausbruch sei „mit Ansage“ gekommen.</td>
<td></td>
</tr>
<tr>
<td>beam (=4)</td>
<td>The outbreak came &quot;with announcement&quot;.</td>
<td>-2.82</td>
</tr>
<tr>
<td>human-A</td>
<td>The outbreak occurred “predictably”.</td>
<td>-18.10</td>
</tr>
<tr>
<td>human-B</td>
<td>The outbreak happened “on cue.”</td>
<td>-18.74</td>
</tr>
</tbody>
</table>

- **Beam search:**
  - Generates target words that are frequent in the training data
  - Translates very literal without considering the sentence context too much

- **Consequences**
  - **Domain mismatch** (generate most likely word-to-word translation)
  - Can introduce **omission** errors (words “hard” to explain)
  - Sounds **awkward and unnatural**

- Human translation have very low pplx as humans use words and sentences structures that are rare in the training data!
Motivation: MAP Decode for Translation Task

<table>
<thead>
<tr>
<th>system</th>
<th>translations</th>
<th>log pplx</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Der Ausbruch sei „mit Ansage“ gekommen.</td>
<td></td>
</tr>
<tr>
<td>beam (=4)</td>
<td>The outbreak came &quot;with announcement&quot;.</td>
<td>-2.82</td>
</tr>
<tr>
<td>model sample</td>
<td>The outbreak happened &quot;with announcement&quot;.</td>
<td>-6.03</td>
</tr>
<tr>
<td>model sample</td>
<td>The outbreak occurred &quot;with announcement&quot;.</td>
<td>-6.61</td>
</tr>
<tr>
<td>model sample</td>
<td>(Table of Contents)</td>
<td>-16.15</td>
</tr>
<tr>
<td>model sample</td>
<td>The outbreak took a &quot;say-so&quot;.</td>
<td>-18.38</td>
</tr>
<tr>
<td>human-A</td>
<td>The outbreak occurred “predictably”.</td>
<td>-18.10</td>
</tr>
<tr>
<td>human-B</td>
<td>The outbreak happened “on cue.”</td>
<td>-18.74</td>
</tr>
</tbody>
</table>

Potential solutions:
- Training data distribution
- Training objective / model
- Inference strategy (in this talk)
  - Instead of the most likely translation (based on the model), we generate the most acceptable translation
Minimum Bayes Risk Decoding

Are we using the right **decoding criterion**?

- Beam search = Maximum LogP ≠ Maximum utility (BLEU, BLEURT, YISI, CHRF...)

MBR decoding

- Take $S$, i.e. $N$ samples from model and find the utility "centroid"

$$h_{MBR} = \arg \max_{h \in S} \frac{1}{N} \sum_{h' \in S} u(h; h')$$

- Need:
  - good model (probability distribution good estimate of $P_{human}(y|x)$)
  - good utility (BLEURT, YISI, CHRF, BLEU?)
Utility Functions

<table>
<thead>
<tr>
<th>Model</th>
<th>log pplx</th>
<th>BLEU</th>
<th>ChrF</th>
<th>YiSi</th>
<th>BLEURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Translation</td>
<td>-38.0</td>
<td>31.5</td>
<td>60.9</td>
<td>84.7</td>
<td>37.1</td>
</tr>
<tr>
<td>Beam (=4)</td>
<td>-11.5</td>
<td>35.2</td>
<td>63.0</td>
<td>85.6</td>
<td>30.3</td>
</tr>
</tbody>
</table>

- **BLEU, ChrF and YiSi**
  - Word/emb-based overlap metrics
  - Aligned with log ppl
  - Idea: explain every single token in the hypotheses with tokens in the reference

- **BLEURT**
  - Projects sentences into an embedding space
  - The sentence structure and the actual token play a secondary role
  - The semantic and the fluency are important
Experimental Setup

- **Language Pairs:**
  - De<->En

- **Training data:**
  - WMT2019 - 57M parallel sentences (paracrawl, nc-v15, europarl, commoncrawl)
  - Filtered via CDS (indomain = nc-v15)

- **Test set:**
  - newstest2019 (dev), newstest2021 (test)

- **Model:**
  - Transformer-big, 300k training steps
  - Model trained w/o label smoothing

- **MBR Decoding:**
  - Sampling strategy: ancestral sampling
  - Candidate Size: 1000
  - Utility function: sentenceBLEU, Chrf, YiSi, BLEURT
## English→German newstest2021 (ref-C)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>sBLEU</th>
<th>Chrf</th>
<th>YiSi</th>
<th>BLEURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Translation (ref-D)</td>
<td>31.5</td>
<td>31.6</td>
<td>60.9</td>
<td>84.7</td>
<td>75.6</td>
</tr>
<tr>
<td>Beam (=4)</td>
<td><strong>34.3</strong></td>
<td>34.2</td>
<td>62.5</td>
<td>85.3</td>
<td>71.6</td>
</tr>
</tbody>
</table>
### English→German newstest2021 (ref-C)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>sBLEU</th>
<th>ChrF</th>
<th>YiSi</th>
<th>BLEURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Translation (ref-D)</td>
<td>31.5</td>
<td>31.6</td>
<td>60.9</td>
<td>84.7</td>
<td>75.6</td>
</tr>
<tr>
<td>Beam (=4)</td>
<td>34.3</td>
<td>34.2</td>
<td>62.5</td>
<td>85.3</td>
<td>71.6</td>
</tr>
<tr>
<td>MBR-sBLEU-add_k(k=1)</td>
<td>34.7</td>
<td><strong>34.8</strong></td>
<td>62.5</td>
<td>85.4</td>
<td>70.5</td>
</tr>
<tr>
<td>MBR-CHRF</td>
<td>34.2</td>
<td>34.3</td>
<td><strong>64.1</strong></td>
<td>85.7</td>
<td>71.4</td>
</tr>
<tr>
<td>MBR-YiSi</td>
<td>34.2</td>
<td>34.2</td>
<td>62.8</td>
<td><strong>86.0</strong></td>
<td>71.6</td>
</tr>
</tbody>
</table>
### English→German newstest2021 (ref-C)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>sBLEU</th>
<th>ChrF</th>
<th>YiSi</th>
<th>BLEURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Translation (ref-D)</td>
<td>31.5</td>
<td>31.6</td>
<td>60.9</td>
<td>84.7</td>
<td>75.6</td>
</tr>
<tr>
<td>Beam (=4)</td>
<td><strong>34.3</strong></td>
<td>34.2</td>
<td>62.5</td>
<td>85.3</td>
<td>71.6</td>
</tr>
<tr>
<td>MBR-sBLEU-add_k(k=1)</td>
<td>34.7</td>
<td><strong>34.8</strong></td>
<td>62.5</td>
<td>85.4</td>
<td>70.5</td>
</tr>
<tr>
<td>MBR-CHRF</td>
<td>34.2</td>
<td>34.3</td>
<td><strong>64.1</strong></td>
<td>85.7</td>
<td>71.4</td>
</tr>
<tr>
<td>MBR-YiSi</td>
<td>34.2</td>
<td>34.2</td>
<td>62.8</td>
<td><strong>86.0</strong></td>
<td>71.6</td>
</tr>
<tr>
<td>MBR-BLEURT</td>
<td>25.4*</td>
<td>26.0</td>
<td>57.7</td>
<td>83.1</td>
<td><strong>79.0</strong></td>
</tr>
</tbody>
</table>

*For more details of the biases of BLEU and why it is good to reduce BLEU scores, read: BLEU might be Guilty but References are not Innocent (EMNLP 2020)*
## English → German newstest2021 (ref-C)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>sBLEU</th>
<th>ChrF</th>
<th>YiSi</th>
<th>BLEURT</th>
<th>log ppl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Translation (ref-D)</td>
<td>31.5</td>
<td>31.6</td>
<td>60.9</td>
<td>84.7</td>
<td>75.6</td>
<td>-38.0</td>
</tr>
<tr>
<td>Beam (=4)</td>
<td><strong>34.3</strong></td>
<td>34.2</td>
<td>62.5</td>
<td>85.3</td>
<td>71.6</td>
<td>-11.5</td>
</tr>
<tr>
<td>MBR-sBLEU-add_k(k=1)</td>
<td>34.7</td>
<td><strong>34.8</strong></td>
<td>62.5</td>
<td>85.4</td>
<td>70.5</td>
<td>-11.2</td>
</tr>
<tr>
<td>MBR-CHRF</td>
<td>34.2</td>
<td>34.3</td>
<td><strong>64.1</strong></td>
<td>85.7</td>
<td>71.4</td>
<td>-13.2</td>
</tr>
<tr>
<td>MBR-YiSi</td>
<td>34.2</td>
<td>34.2</td>
<td>62.8</td>
<td><strong>86.0</strong></td>
<td>71.6</td>
<td>-11.4</td>
</tr>
<tr>
<td>MBR-BLEURT</td>
<td>25.4</td>
<td>26.0</td>
<td>57.7</td>
<td>83.1</td>
<td><strong>79.0</strong></td>
<td>-24.4</td>
</tr>
</tbody>
</table>
## English→German newstest2021 (ref-C)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>sBLEU</th>
<th>ChrF</th>
<th>YiSi</th>
<th>BLEURT</th>
<th>log ppl</th>
<th>MQM human eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Translation (ref-D)</td>
<td>31.5</td>
<td>31.6</td>
<td>60.9</td>
<td>84.7</td>
<td>75.6</td>
<td>-38.0</td>
<td>0.338</td>
</tr>
<tr>
<td>Beam (=4)</td>
<td><strong>34.3</strong></td>
<td>34.2</td>
<td>62.5</td>
<td>85.3</td>
<td>71.6</td>
<td>-11.5</td>
<td>2.392</td>
</tr>
<tr>
<td>MBR-sBLEU-add_k(k=1)</td>
<td>34.7</td>
<td><strong>34.8</strong></td>
<td>62.5</td>
<td>85.4</td>
<td>70.5</td>
<td>-11.2</td>
<td><strong>1.992</strong></td>
</tr>
<tr>
<td>MBR-CHRF</td>
<td>34.2</td>
<td>34.3</td>
<td><strong>64.1</strong></td>
<td>85.7</td>
<td>71.4</td>
<td>-13.2</td>
<td>2.214</td>
</tr>
<tr>
<td>MBR-YiSi</td>
<td>34.2</td>
<td>34.2</td>
<td>62.8</td>
<td><strong>86.0</strong></td>
<td>71.6</td>
<td>-11.4</td>
<td>2.842</td>
</tr>
<tr>
<td>MBR-BLEURT</td>
<td>25.4</td>
<td>26.0</td>
<td>57.7</td>
<td>83.1</td>
<td><strong>79.0</strong></td>
<td>-24.4</td>
<td>1.562</td>
</tr>
</tbody>
</table>

1. **MBR works:** Each MBR-utility is best on their utility
2. **Human evaluation**
   a. **MBR-BLEU** outperforms beam search decoding
   b. **MBR-BLEURT** wins by a huge margin
## MQM - Human Evaluation with Error Categories

### Error Category

<table>
<thead>
<tr>
<th>Error Category</th>
<th>beam</th>
<th>MBR</th>
<th>BLEURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy/Omission</td>
<td>18</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Terminology/Inappropriate for context</td>
<td>151</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Style/Awkward</td>
<td>66</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Accuracy/Mistranslation</td>
<td>70</td>
<td>58</td>
<td></td>
</tr>
</tbody>
</table>

Number of major errors

### Improvements:

1. **Beam Search Problem:**
   - Length of translations
   - Example: omission

2. **Beam Search Problem:**
   - Autoregressive decoding
   - Example: inappropriate for context

3. **Beam Search Problem:**
   - Generate most probably tokens
   - Example 1: Style/Awkward
   - Example 2: Mistranslation
## Example Translations

<table>
<thead>
<tr>
<th>System</th>
<th>translations</th>
<th>log pplx</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Der Ausbruch sei „mit Ansage“ gekommen.</td>
<td></td>
</tr>
<tr>
<td>beam (=4)</td>
<td>The outbreak <em>came</em> &quot;with announcement&quot;.</td>
<td>-2.82</td>
</tr>
<tr>
<td>MBR-sBLEU</td>
<td>The outbreak <em>came</em> &quot;with announcement&quot;.</td>
<td>-2.82</td>
</tr>
<tr>
<td>MBR-Chrf</td>
<td>The outbreak <em>came</em> &quot;with announcement&quot;.</td>
<td>-2.82</td>
</tr>
<tr>
<td>MBR-YiSi</td>
<td>The outbreak <em>came</em> &quot;with announcement&quot;.</td>
<td>-2.82</td>
</tr>
<tr>
<td>MBR-BLEURT</td>
<td>The outbreak <em>occurred</em> &quot;as announced&quot;.</td>
<td>-11.21</td>
</tr>
<tr>
<td>human-A</td>
<td>The outbreak <em>occurred</em> “predictably”.</td>
<td>-18.10</td>
</tr>
<tr>
<td>human-B</td>
<td>The outbreak <em>happened</em> “on cue.”</td>
<td>-18.74</td>
</tr>
</tbody>
</table>

- MBR with sBLEU, chrf, or YiSi generate the same translation as beam search decoding
- MBR-BLEURT generates a better, more natural translation
No Beam Search Curse

- Quality improves when scaling the number of candidates
- Overcoming typical problems with beam search decoding
Pruning

- Randomly sampling on the expectation side is promising
- Caveat: mean BLEURT scores hide a lot of information
MBR Generates Different Translations

- MBR-BLEURT translations are quite different compared to beam search and MBR with overlap metrics
- Similarly different like 2 different human evaluation
Conclusions

1. MBR decoding with a neural metric like BLEURT significantly outperforms beam search decoding.
2. MBR decoding with BLEU outperforms beam search decoding.
3. MBR-BLEURT overcomes many problems of beam search decoding:
   - Omissions errors
   - Mistranslation
   - Awkward style
   - “Beam search curse”
4. MBR-BLEURT generates translations with much lower model probabilities:
   - More similar to the style of human translations
5. Many more experiments in our paper:
   - More language pairs
   - Impact of different NMT models
   - Pruning strategies
   - ...
Findings & Recommendations

1. Crowd-worker human evaluation has low correlation with MQM
   a. Unable to distinguish MT from human translation
   b. Difference in ranking of WMT submissions
   c. Bad ranking of online systems
   d. Same finding for WMT21 (check out Sec 8.3 of the Metric Task paper)

2. Stop using crowd-worker for human evals
   a. Unreliable, biased
   b. We have experiments comparing different rater pools based on the same human eval in the paper

3. Use MQM
   a. Reliable evaluation also for closer systems
   b. Error Annotation will help to understand the difference between 2 systems
   c. Error annotations should guide MT research
   d. Flexible error weighting schema

4. Higher correlation of automatic metrics with MQM
   a. WMT human eval correlation with MQM lower than most of the metrics
   b. MQM annotations are extremely helpful to improve and evaluate automatic metrics
      i. WMT21 and WMT22 Metric Task are using them

5. All data is available on github:
   b. MQM viewer: https://github.com/google-research/google-research/tree/master/mqm_viewer
Future Research Directions

**Human Evaluation:**
1. Establish MQM as a standard for human evaluation
   a. Make MQM and the annotators accessible to everyone
2. Improve Inter-annotator agreement of raters so that we can compare MQM evaluations from different rater pools
3. Invent new human evaluation methodologies that have high correlation with MQM, but cheaper
   a. Is it possible to do this with crowd workers?
4. Use MQM in other NLP fields
   a. We expect similar findings

**Automatic Evaluation:**
5. Develop trained metrics that can predict error spans and error categories
   a. Help understand errors from systems
   b. Make MQM more accessible to everyone

**MT modelling research:**
6. Develop new interesting approaches
   a. E.g. Paragraph-level translations
7. Re-visit existing approaches that were under-evaluated before?
   a. Pre-training/ LM-augmented NMT
References

For papers discussed in this talk

1. APE at scale and its implications on MT evaluation biases - M Freitag, I Caswell, S Roy [WMT 19]
2. Translationese as a Language in" Multilingual" NMT - P Riley, I Caswell, M Freitag, D Grangier [ACL 19]
3. BLEU might be guilty but references are not innocent - M Freitag, D Grangier, I Caswell [EMNLP 20]
4. Results of the WMT20 metrics shared task - N Mathur, J Wei, M Freitag, Q Ma, O Bojar [WMT 20]
5. Experts, errors, and context: A large-scale study of human evaluation for machine translation - M Freitag, G Foster, D Grangier, V Ratnakar, Q Tan, W Macherey [TACL 21]
6. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain - M Freitag, R Rei, N Mathur, C Lo, C Stewart, G Foster, A Lavie, O Bojar [WMT 21]
7. Minimum Bayes Risk Decoding with Neural Metrics of Translation Quality - M Freitag, D Grangier, Q Tan, B Liang [TACL 22]
8. A Natural Diet: Towards Improving Naturalness of Machine Translation Output - M Freitag, D Vilar, D Grangier, C Cherry, G Foster [ACL 22]
Thank you