Outline

1. Quality evaluation
2. Reference-based metrics
3. Quality estimation metrics
4. Metrics in the NMT era
Why do we care?

... or why is this the first lecture of the Marathon?

In the business of developing MT, we need to

- measure progress over new/alternative versions
- compare different MT systems
- decide whether a translation is good enough for something
- optimise parameters of MT systems
- understand where systems go wrong (diagnosis)
Why do we care?

- One should optimise a system using the same metric that will be used to evaluate it.
- **Issue**: how to choose a metric? Choice should be related to the purpose of the system will be used (not the case in practice).
- Other aspects are important for tuning (sentence/corpus-level, fast, cheap, differentiable, ...).
“MT evaluation is better understood than MT”
(Carbonell and Wilks, 1991)
“MT evaluation is better understood than MT”
(Carbonell and Wilks, 1991)
“MT evaluation is better understood than MT”
   (Carbonell and Wilks, 1991)

“There are more MT evaluation metrics than MT approaches”
   (Specia, 2016)
**Complex problem**

- What does **quality** mean?
  - Fluent? Adequate? Both?
  - Easy to post-edit?
  - System A better than system B?
  - ...
Complex problem

- What does quality mean?
  - Fluent? Adequate? Both?
  - Easy to post-edit?
  - System A better than system B?
  - ...

- Quality for whom/what?
  - End-user (gisting vs dissemination)
  - Post-editor (light vs heavy post-editing)
  - Other applications (e.g. CLIR)
  - MT-system (tuning or diagnosis for improvement)
  - ...
Complex problem

**MT**  Do buy this product, it’s their craziest invention!
Complex problem

MT  Do buy this product, it’s their craziest invention!
HT  Do **not** buy this product, it’s their craziest invention!
Complex problem

MT  Do buy this product, it’s their craziest invention!

HT  Do **not** buy this product, it’s their craziest invention!

- **Severe** if end-user does not speak source language
- **Trivial** to post-edit by translators
Complex problem

MT Six-hours battery, 30 minutes to full charge last.
Complex problem

MT  Six-hours battery, 30 minutes to full charge last.

HT  The battery lasts 6 hours and it can be fully recharged in 30 minutes.
MT  Six-hours battery, **30 minutes** to **full charge** last.

HT  The **battery lasts** 6 hours and it can be **fully recharged** in **30 minutes**.

- **Ok** for gisting - meaning preserved
- **Very costly** for post-editing if style is to be preserved
A taxonomy of MT evaluation methods

- Manual
- Automatic
A taxonomy of MT evaluation methods

Manual

Automatic

Direct asses.

Scoring
A taxonomy of MT evaluation methods

**Source:** les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l’UE.

**Reference:** rather, the two countries form a laboratory needed for the internal working of the EU.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>both countries are rather a necessary laboratory the internal operation of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>

Is this translation correct? 

---

Read the text below. How much do you agree with the following statement:

The black text adequately expresses the meaning of the gray text in English.

---

To snobs like me who declare that they’d rather play sports than watch them, it’s hard to see the appeal of watching games rather than taking up a controller myself.

Snob like me, who say that it is better to be in sports than watching him, it is hard to understand the appeal of having to watch the game, rather than to take a joystick in hand.

---

0 % 100 %
A taxonomy of MT evaluation methods

Manual

Direct asses.

Automatic

Scoring

Ranking

Translation Quality Assessment: Evaluation and Estimation
A taxonomy of MT evaluation methods

<table>
<thead>
<tr>
<th>Appraise</th>
<th>Overview</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Până la mijlocul lui iulie, procentul a urcat la 40%. La începutul lui august, era 52%.</td>
<td>— Source</td>
<td></td>
</tr>
<tr>
<td>By mid-July, it was 40 percent. In early August, it was 52 percent.</td>
<td>— Reference</td>
<td></td>
</tr>
</tbody>
</table>

- Best ← Rank 1 ← Rank 2 ← Rank 3 ← Rank 4 ← Rank 5 ← Worst
- Until the middle of July, the percentage rose to 40%.
- Until mid-July, the percentage rose to 40%.
- By mid-July, the percentage climbed to 40 per cent.
- Until mid-July, the percentage climbed to 40%.
- Until the middle of July, the figure climbed to 40%.
A taxonomy of MT evaluation methods

- **Manual**
  - Direct assessment
  - Scoring
  - Ranking
  - Error annotation

- **Automatic**
A taxonomy of MT evaluation methods
A taxonomy of MT evaluation methods

Manual
- Direct assessment
- Scoring
- Ranking
- Error annotation
- Post-editing

Task-based

Automatic
### A taxonomy of MT evaluation methods

<table>
<thead>
<tr>
<th>Quality evaluation</th>
<th>Reference-based metrics</th>
<th>Quality estimation metrics</th>
<th>Metrics in the NMT era</th>
</tr>
</thead>
</table>

**In patients with ability of hypersensitivity (allergy) to manganolipol or any of the other ingredients of Tisalcain must not be used.**

<table>
<thead>
<tr>
<th>Translation Quality Assessment: Evaluation and Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTER</td>
</tr>
</tbody>
</table>

- This table lists various methods for evaluating machine translation (MT) quality, including reference-based metrics and quality estimation metrics.

- The table contains translations from German to English, indicating the use of MT for communication purposes.

- The heading of the table is "A taxonomy of MT evaluation methods."
A taxonomy of MT evaluation methods

- **Seconds per word**
- **HTER**
A taxonomy of MT evaluation methods

Manual
- Direct assessment
  - Scoring
  - Ranking
  - Error annotation

Task-based
- Post-editing
- Reading comprehension

Automatic
A taxonomy of MT evaluation methods

In recent years, there has been a growing movement to "retire" the penny or take it out of circulation. This movement has been countered by people passionate about preserving the penny. There are compelling reasons to eliminate the penny and to preserve it. What do you think?

ELIMINATE THE PENNY

According to the U.S. Mint, it costs 2.4 cents to produce one penny. In other words, the cost of making a penny is more than double its value. Since the United States Mint produced $50 million worth of pennies in 2010 at a cost of $120 million dollars, $70 million was wasted.

Advocates of "retiring" the penny claim the coin is obsolete and virtually worthless. Nothing can realistically be bought for a penny anymore. In addition, simply handling pennies...
A taxonomy of MT evaluation methods

Manual
- Direct assessment
  - Scoring
  - Ranking
  - Error annotation

Task-based
- Post-editing
- Reading comprehension
- Eye-tracking

Automatic
A taxonomy of MT evaluation methods
A taxonomy of MT evaluation methods

Manual
- Direct assessment
  - Scoring
  - Ranking
  - Error annotation

Task-based
- Post-editing
- Reading comprehension
- Eye-tracking

Automatic
- Reference-based
A taxonomy of MT evaluation methods

Manual
- Direct assess.
- Task-based
  - Scoring
  - Ranking
  - Error annotation
  - Post-editing
  - Reading comprehension
  - Eye-tracking

Automatic
- Reference-based
- Quality estimation

Reference-based metrics
- BLEU, Meteor, NIST, TER, WER, PER, CDER, BEER, CiDER, Cobalt, RATATOUILLE, RED, AMBER, PARMESAN, ...

Quality estimation metrics

Metrics in the NMT era

Translation Quality Assessment: Evaluation and Estimation
A taxonomy of MT evaluation methods

Manual
- Direct assessment
  - Scoring
  - Ranking
- Task-based
  - Error annotation
  - Post-editing
  - Reading comprehension
  - Eye-tracking

Automatic
- Reference-based
- Quality estimation

Reference-based metrics
- BLEU, Meteor, NIST, TER, WER, PER, CDER, BEER, CIDER, Cobalt, RATATOUILLE, RED, AMBER, PARMESAN, ...

Quality estimation metrics
A taxonomy of MT evaluation methods

- **Manual**
  - Direct assessment
  - Task-based
    - Scoring
    - Ranking
    - Error annotation
    - Post-editing
    - Reading comprehension
    - Eye-tracking

- **Automatic**
  - Reference-based
    - BLEU, Meteor, NIST, TER, WER, PER, CDER, BEER, CiDER, Cobalt, RATATOUILLE, RED, AMBER, PARMESAN, ...
  - Quality estimation
Outline

1. Quality evaluation
2. Reference-based metrics
3. Quality estimation metrics
4. Metrics in the NMT era
Assumption

The closer an MT system output is to a human translation (HT = reference), the better it is.

Which system is better?

\[ \text{MT}_1 \] Indignation in front of photos of a veiled woman controlled on the beach in Nice.

\[ \text{MT}_2 \] Outrage at pictures of a veiled woman controlled on the beach in Nice.

\[ \text{HT}_a \] Indignation at pictures of a veiled woman being checked on a beach in Nice.
The closer an MT system output is to a human translation (HT = reference), the better it is.

Which system is better?

**MT₁**  Indignation in front of photos of a veiled woman controlled on the beach in Nice.

**MT₂**  Outrage at pictures of a veiled woman controlled on the beach in Nice.

**HTₐ**  Indignation at pictures of a veiled woman being checked on a beach in Nice.

Or, simply, how good is the MT₁ system output?
Assumption

Which system is better?

**MT_1**  Indignation in front of photos of a veiled woman controlled on the beach in Nice.

**MT_2**  Outrage at pictures of a veiled woman controlled on the beach in Nice.

**HT_a**  Indignation at pictures of a veiled woman being checked on a beach in Nice.

**HT_b**  Photos of a veiled woman checked by the police on the beach in Nice cause outrage.
Assumption

Which system is better?

\textbf{MT}_1 \quad \text{Indignation in front of photos of a veiled woman controlled on the beach in Nice.}

\textbf{MT}_2 \quad \text{Outrage at pictures of a veiled woman controlled on the beach in Nice.}

\textbf{HT}_a \quad \text{Indignation at pictures of a veiled woman being checked on a beach in Nice.}

\textbf{HT}_b \quad \text{Photos of a veiled woman checked by the police \textbf{on the beach in Nice} cause \textbf{outrage}.}

Or, again, how good is the \textbf{MT}_1 system output?
BLEU: BiLingual Evaluation Understudy

- **Most widely used metric**, both for MT system evaluation/comparison and SMT tuning
- Matching of n-grams between MT and HT: rewards same words in equal order
- \(#clip(g)\) count of reference n-grams \(g\) which happen in a MT sentence \(h\) clipped by the number of times \(g\) appears in the HT sentence for \(h\); \(\#(g') = \) number of n-grams in MT output
- n-gram precision \(p_n\) for a set of translations in \(C\):

\[
p_n = \frac{\sum_{c \in C} \sum_{g \in \text{ngrams}(c)} \#clip(g)}{\sum_{c \in C} \sum_{g' \in \text{ngrams}(c)} \#(g')}
\]
BLEU

- Combine (mean of the log) 1-n n-gram precisions

\[ \sum_{n} \log p_n \]
BLEU

- Combine (mean of the log) 1-n n-gram precisions
  \[ \sum_n \log p_n \]

- Bias towards translations with fewer words
- **Brevity penalty** to penalise MT sentences that are shorter than reference
  - Compares the overall number of words $w_h$ of the entire hypotheses set with ref length $w_r$:
    \[ BP = \begin{cases} 
    1 & \text{if } w_c \geq w_r \\
    e^{(1-w_r/w_c)} & \text{otherwise}
    \end{cases} \]
BLEU

- Combine (mean of the log) 1-n n-gram precisions
  \[ \sum_n \log p_n \]

- Bias towards translations with fewer words
- **Brevity penalty** to penalise MT sentences that are shorter than reference
  - Compares the overall number of words \( w_h \) of the entire hypotheses set with ref length \( w_r \):
    \[
    BP = \begin{cases} 
    1 & \text{if } w_c \geq w_r \\ 
    e^{(1-w_r/w_c)} & \text{otherwise} 
    \end{cases}
    \]

  \[ \text{BLEU} = BP \times \exp \left( \sum_n \log p_n \right) \]
BLEU

- Scale: 0-1, but highly **dependent on the test set**
- Rewards **fluency** by matching high n-grams (up to 4)
- Rewards **adequacy** by unigrams and brevity penalty – poor model of recall
- **Synonyms and paraphrases** only handled if in one of reference translations
- All tokens are **equally weighted**: incorrect content word = incorrect determiner
BLEU

- Scale: 0-1, but highly dependent on the test set
- Rewards **fluency** by matching high n-grams (up to 4)
- Rewards **adequacy** by unigrams and brevity penalty – poor model of recall
- **Synonyms and paraphrases** only handled if in one of reference translations
- All tokens are **equally weighted**: incorrect content word = incorrect determiner
- Better for evaluating changes in the same system than comparing **different MT architectures**
Example:

**MT:** in two weeks Iraq’s weapons will give army

**HT:** the Iraqi weapons are to be handed over to the army within two weeks

- 1-gram precision: 4/8
- 2-gram precision: 1/7
- 3-gram precision: 0/6
- 4-gram precision: 0/5
Edit distance metrics

**TER: Translation Error Rate**

- Levenshtein edit distance
- Minimum proportion of *insertions*, *deletions*, and *substitutions* to transform MT sentence into HT
- Adds *shift* operation

**Example:**

REF: SAUDI ARABIA denied this week information published in the AMERICAN new york times

HYP: [this week] the saudis denied information published in the ***** new york times

1 shift, 2 substit., 1 deletion: TER = \( \frac{4}{13} \approx 0.31 \)
Edit distance metrics

**TER: Translation Error Rate**
- Levenshtein edit distance
- Minimum proportion of *insertions, deletions, and substitutions* to transform MT sentence into HT
- Adds *shift* operation

REF: *SAUDI ARABIA* denied this week information published in the *AMERICAN* new york times

HYP: [this week] the saudis denied information published in the ***** new york times

1 shift, 2 substit., 1 deletion: TER = \( \frac{1}{3} \).
Edit distance metrics

**TER**: Translation Error Rate
- Levenshtein edit distance
- Minimum proportion of *insertions, deletions, and substitutions* to transform MT sentence into HT
- Adds *shift* operation

REF: SAUDI ARABIA denied this week information published in the AMERICAN new york times
HYP: [this week] the saudis denied information published in the ***** new york times

1 shift, 2 substit., 1 deletion: \( \text{TER} = \frac{4}{13} = 0.31 \)
Edit distance metrics

**TER: Translation Error Rate**

- Levenshtein edit distance
- Minimum proportion of *insertions, deletions, and substitutions* to transform MT sentence into HT
- Adds *shift* operation

REF: **SAUDI ARABIA** denied this week information published in the **AMERICAN** new york times

HYP: [this week] the saudis denied information published in the ***** new york times

1 shift, 2 substit., 1 deletion: \( \text{TER} = \frac{4}{13} = 0.31 \)

**Human-targeted TER (HTER)**

TER between MT and its post-edited version
Alignment-based metrics

**METEOR:**

- **Unigram** Precision and Recall
- Align MT & HT
- Matching considers *inflection variants* (stems), **synonyms**, **paraphrases**
- **Fluency** addressed via a direct penalty: fragmentation of the matching
- METEOR score = F-mean score discounted for fragmentation = F-mean * (1 - DF)
- Parameters can be trained
Alignment-based metrics

**MT:** in two weeks Iraq’s weapons will give army

**HT:** the Iraqi weapons are to be handed over to the army within two weeks
Alignment-based metrics

MT: in **two weeks** Iraq’s weapons will give army

HT: the **Iraqi weapons** are to be handed over to the army within **two weeks**

- **Matching:**

  MT: **two weeks Iraq’s weapons army**

  HT: Iraqi weapons army **two weeks**

\[
P = \frac{5}{8} = 0.625 \\
R = \frac{5}{14} = 0.357 \\
F\text{-mean} = \frac{10 \times P \times R}{9P + R} = 0.373 \\
\text{Fragmentation: } \frac{3}{5} = 0.6 \\
\text{Discounting factor: } DF = 0.5 \times (0.6)^3 = 0.108 \\
\text{METEOR: } F\text{-mean} \times (1 - DF) = 0.373 \times 0.892 = 0.333\]
Alignment-based metrics

**MT:** in *two weeks* *Iraq’s weapons* will give *army*

**HT:** the *Iraqi weapons* are to be handed over to the *army* within *two weeks*

- Matching:

  **MT** *two weeks Iraq’s weapons army*
  
  **HT:** *Iraqi weapons army two weeks*

- \( P = \frac{5}{8} = 0.625 \)
- \( R = \frac{5}{14} = 0.357 \)
- \( F\text{-mean} = \frac{10 \times P \times R}{9P + R} = 0.373 \)
Alignment-based metrics

MT: in two weeks Iraq’s weapons will give army
HT: the Iraqi weapons are to be handed over to the army within two weeks
  - Matching:
MT two weeks Iraq’s weapons army
HT: Iraqi weapons army two weeks
  - P = 5/8 = 0.625
  - R = 5/14 = 0.357
  - F-mean = 10*P*R/(9P+R) = 0.373
  - Fragmentation: 3 frags for 5 words = (3)/(5) = 0.6
  - Discounting factor: DF = 0.5 * (0.6**3) = 0.108
  - METEOR: F-mean * (1 - DF) = 0.373 * 0.892 = 0.333
BEER: BEtter Evaluation as Ranking

- **Trained metric**
  \[
  \text{score}(h, r) = \sum_i w_i \times \phi_i(h, r) = \mathbf{w} \cdot \mathbf{\phi}
  \]
- Learns from pairwise rankings
- Various features between MT output and reference translation
  - Precision, Recall and F1 over character n-grams (1-6)
  - Idem for word unigrams: content vs function separately
  - Reordering through permutation trees and distance to ideal monotone permutation
Dozens more....

**Some - WMT metrics task:**

- CharacTer
- chrF/wordF
- TerroCat
- MEANT and TINE
- TESLA
- LEPOR
- ROSE
- AMBER

- Many other linguistically motivated metrics where matching goes beyond word forms
- ...

**Asiya toolkit - up until ~2014**
WMT16 metrics task (by Bojar et al.):

<table>
<thead>
<tr>
<th>Metric</th>
<th># Wins</th>
<th>Language Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEER</td>
<td>11</td>
<td>csen, encs, ende, enfi, enro, enru, entr, fien, roen, ruen, tren</td>
</tr>
<tr>
<td>UoW.ReVal</td>
<td>6</td>
<td>csen, deen, fien, roen, ruen, tren</td>
</tr>
<tr>
<td>chrF2</td>
<td>6</td>
<td>csen, encs, enro, entr, fien, ruen</td>
</tr>
<tr>
<td>chrF1</td>
<td>5</td>
<td>encs, enro, fien, ruen, tren</td>
</tr>
<tr>
<td>chrF3</td>
<td>4</td>
<td>deen, enfi, entr, ruen</td>
</tr>
<tr>
<td>mosesCDER</td>
<td>4</td>
<td>csen, enfi, enru, entr</td>
</tr>
<tr>
<td>CharacTer</td>
<td>3</td>
<td>csen, deen, roen</td>
</tr>
<tr>
<td>mosesBLEU</td>
<td>3</td>
<td>csen, encs, enfi</td>
</tr>
<tr>
<td>mosesPER</td>
<td>3</td>
<td>enro, ruen, tren</td>
</tr>
<tr>
<td>mtevalBLEU</td>
<td>3</td>
<td>csen, encs, enro</td>
</tr>
<tr>
<td>wordF1</td>
<td>3</td>
<td>csen, encs, enro</td>
</tr>
<tr>
<td>wordF2</td>
<td>3</td>
<td>csen, encs, enro</td>
</tr>
<tr>
<td>mosesTER</td>
<td>2</td>
<td>csen, encs</td>
</tr>
<tr>
<td>mtevalNIST</td>
<td>2</td>
<td>encs, tren</td>
</tr>
<tr>
<td>wordF3</td>
<td>2</td>
<td>csen, entr</td>
</tr>
<tr>
<td>mosesWER</td>
<td>1</td>
<td>csen</td>
</tr>
</tbody>
</table>
Problems with reference-based evaluation

- Reference(s): subset of good translations, usually **one**
  - Some metrics expand matching, e.g. synonyms in **Meteor**

- Huge **variation** in reference translations. E.g.

<table>
<thead>
<tr>
<th>Source</th>
<th>不过这一切都由不得你</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT</td>
<td>But all this is beyond the control of you.</td>
</tr>
<tr>
<td>HT&lt;sub&gt;1&lt;/sub&gt;</td>
<td>But all this is beyond your control.</td>
</tr>
<tr>
<td>HT&lt;sub&gt;2&lt;/sub&gt;</td>
<td>However, you cannot choose yourself.</td>
</tr>
<tr>
<td>HT&lt;sub&gt;3&lt;/sub&gt;</td>
<td>However, not everything is up to you to decide.</td>
</tr>
<tr>
<td>HT&lt;sub&gt;4&lt;/sub&gt;</td>
<td>But you can’t choose that.</td>
</tr>
</tbody>
</table>

- **Human score** | **BLEU score**

<table>
<thead>
<tr>
<th></th>
<th>3.4</th>
<th>0.427</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT&lt;sub&gt;1&lt;/sub&gt;</td>
<td>2</td>
<td>0.049</td>
</tr>
<tr>
<td>HT&lt;sub&gt;2&lt;/sub&gt;</td>
<td>2</td>
<td>0.050</td>
</tr>
<tr>
<td>HT&lt;sub&gt;3&lt;/sub&gt;</td>
<td>2.8</td>
<td>0.055</td>
</tr>
</tbody>
</table>

- Metrics completely disregard **source segment**

- **Cannot** be applied for MT systems in use
Problems with reference-based evaluation

- **Reference(s):** subset of good translations, usually **one**
  - Some metrics expand matching, e.g. synonyms in **Meteor**

- **Huge variation** in reference translations. E.g.

<table>
<thead>
<tr>
<th>Source</th>
<th>不过这一切都由不得你</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT</td>
<td>But all this is beyond the control of you.</td>
</tr>
<tr>
<td>HT₁</td>
<td>But all this is beyond your control.</td>
</tr>
<tr>
<td>HT₂</td>
<td>However, you cannot choose yourself.</td>
</tr>
<tr>
<td>HT₃</td>
<td>However, not everything is up to you to decide.</td>
</tr>
<tr>
<td>HT₄</td>
<td>But you can’t choose that.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human score</th>
<th>3.4</th>
<th>0.427</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU score</td>
<td>2</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>2.8</td>
<td>0.055</td>
</tr>
</tbody>
</table>

- **Metrics completely disregard source segment**
- **Cannot** be applied for MT systems in use
Problems with reference-based evaluation

- **Reference(s):** subset of good translations, usually **one**
  Some metrics expand matching, e.g. synonyms in ** Meteor**

- **Huge variation** in reference translations. E.g.

<table>
<thead>
<tr>
<th>Source</th>
<th>不过这一切都由不得你</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MT</strong></td>
<td>But all this is beyond the control of you.</td>
</tr>
<tr>
<td><strong>HT_1</strong></td>
<td>But all this is beyond your control.</td>
</tr>
<tr>
<td><strong>HT_2</strong></td>
<td>However, you cannot choose yourself.</td>
</tr>
<tr>
<td><strong>HT_3</strong></td>
<td>However, not everything is up to you to decide.</td>
</tr>
<tr>
<td><strong>HT_4</strong></td>
<td>But you can’t choose that.</td>
</tr>
<tr>
<td><strong>Human score</strong></td>
<td><strong>BLEU score</strong></td>
</tr>
<tr>
<td>3.4</td>
<td>0.427</td>
</tr>
<tr>
<td>2</td>
<td>0.049</td>
</tr>
<tr>
<td>2</td>
<td>0.050</td>
</tr>
<tr>
<td>2.8</td>
<td>0.055</td>
</tr>
</tbody>
</table>

- **Metrics completely disregard source segment**
- **Cannot** be applied for MT systems in use
Outline

1. Quality evaluation
2. Reference-based metrics
3. Quality estimation metrics
4. Metrics in the NMT era
Quality estimation (QE): metrics that provide an estimate on the quality of translations on the fly
**QE - Overview**

- **Quality estimation (QE):** metrics that provide an estimate on the quality of translations *on the fly*
- Quality defined by the data: **purpose** is clear, no comparison to **references**, **source** considered
Quality estimation (QE): metrics that provide an estimate on the quality of translations on the fly

Quality defined by the data: purpose is clear, no comparison to references, source considered

Quality = Can we publish it as is?
Quality estimation (QE): metrics that provide an estimate on the quality of translations on the fly

Quality defined by the data: purpose is clear, no comparison to references, source considered

- Quality = Can we publish it as is?
- Quality = Can a reader get the gist?
Quality estimation (QE): metrics that provide an estimate on the quality of translations on the fly

Quality defined by the data: purpose is clear, no comparison to references, source considered

Quality = Can we publish it as is?

Quality = Can a reader get the gist?

Quality = Is it worth post-editing it?
Quality estimation (QE): metrics that provide an estimate on the quality of translations on the fly.

Quality defined by the data: purpose is clear, no comparison to references, source considered.

- **Quality** = Can we publish it as is?
- **Quality** = Can a reader get the gist?
- **Quality** = Is it worth post-editing it?
- **Quality** = How much effort to fix it?
QE - Framework

Building a model:

X: examples of source & translations

Feature extraction

Features

Y: Quality scores for examples in X

Machine Learning

QE model
QE - Framework

Applying the model:

MT system → Translation for $x_t$

Source Text $x_s'$ → Feature extraction

Features → Quality score $y'$

QE model
Data and levels of granularity

- **Sentence level**: 1-5 subjective scores, PE time, PE edits
- **Word level**: good/bad, good/delete/replace, MQM
- **Phrase level**: good/bad
- **Document level**: PE effort
**Features and algorithms**

Algorithms can be used off-the-shelf
**QE - baseline setting**

**Features:**

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of seen source n-grams
QE - baseline setting

**Features:**

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of seen source n-grams

**SVM regression** with RBF kernel
Quality evaluation

Reference-based metrics

Quality estimation metrics

Metrics in the NMT era

**QE - baseline setting**

**Features:**

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of seen source n-grams

**SVM regression** with RBF kernel

**QuEst:** http://www.quest.dcs.shef.ac.uk/
## QE - SoA sentence-level

### Predicting HTER (WMT16)

<table>
<thead>
<tr>
<th>System ID</th>
<th>Pearson ↑</th>
<th>Spearman ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English-German</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• YSDA/SNTX+BLEU+SVM</td>
<td>0.525</td>
<td>–</td>
</tr>
<tr>
<td>POSTECH/SENT-RNN-QV2</td>
<td>0.460</td>
<td>0.483</td>
</tr>
<tr>
<td>SHEF-LIUM/SVM-NN-emb-QuEst</td>
<td>0.451</td>
<td>0.474</td>
</tr>
<tr>
<td>POSTECH/SENT-RNN-QV3</td>
<td>0.447</td>
<td>0.466</td>
</tr>
<tr>
<td>SHEF-LIUM/SVM-NN-both-emb</td>
<td>0.430</td>
<td>0.452</td>
</tr>
<tr>
<td>UGENT-LT3/SCATE-SVM2</td>
<td>0.412</td>
<td>0.418</td>
</tr>
<tr>
<td>UFAL/MULTIVEC</td>
<td>0.377</td>
<td>0.410</td>
</tr>
<tr>
<td>RTM/RTM-FS-SVR</td>
<td>0.376</td>
<td>0.400</td>
</tr>
<tr>
<td>UU/UU-SVM</td>
<td>0.370</td>
<td>0.405</td>
</tr>
<tr>
<td>UGENT-LT3/SCATE-SVM1</td>
<td>0.363</td>
<td>0.375</td>
</tr>
<tr>
<td>RTM/RTM-SVR</td>
<td>0.358</td>
<td>0.384</td>
</tr>
<tr>
<td><strong>Baseline SVM</strong></td>
<td>0.351</td>
<td>0.390</td>
</tr>
<tr>
<td>SHEF/SimpleNets-SRC</td>
<td>0.182</td>
<td>–</td>
</tr>
<tr>
<td>SHEF/SimpleNets-TGT</td>
<td>0.182</td>
<td>–</td>
</tr>
</tbody>
</table>

Translation Quality Assessment: Evaluation and Estimation
Outline

1. Quality evaluation
2. Reference-based metrics
3. Quality estimation metrics
4. Metrics in the NMT era
SMT vs NMT

Pearson correlation with DA scores for popular metrics on 200 sentences from WMT16’s **uedin** SMT and NMT systems:

<table>
<thead>
<tr>
<th>Metric</th>
<th>uedin-pbmt</th>
<th>uedin-nmt</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.4433</td>
<td>0.5126</td>
</tr>
<tr>
<td>Meteor</td>
<td>0.5123</td>
<td>0.5781</td>
</tr>
<tr>
<td>TER</td>
<td>-0.4042</td>
<td>-0.5592</td>
</tr>
<tr>
<td>chrF2</td>
<td>0.4959</td>
<td>0.5826</td>
</tr>
<tr>
<td>BEER</td>
<td>0.5034</td>
<td>0.6140</td>
</tr>
<tr>
<td>UPF-Cobalt</td>
<td>0.5365</td>
<td>0.5511</td>
</tr>
<tr>
<td>CobaltF-comp</td>
<td>0.5306</td>
<td>0.6064</td>
</tr>
<tr>
<td>DPMFcomb</td>
<td>0.5757</td>
<td>0.6507</td>
</tr>
</tbody>
</table>

(Work with **Marina Fomicheva**)
Are metrics better for NMT because systems are better?

Correlation with DA scores on 840 **low-quality** (Q1-human) & 840 **high-quality** (Q4-human) sentences (all systems)

<table>
<thead>
<tr>
<th></th>
<th>Q1 - low quality</th>
<th>Q4 - high quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.0338</td>
<td>0.4561</td>
</tr>
<tr>
<td>Meteor</td>
<td>0.1985</td>
<td>0.5143</td>
</tr>
<tr>
<td>TER</td>
<td>-0.0870</td>
<td>-0.3710</td>
</tr>
<tr>
<td>UPF-Cobalt</td>
<td>0.1499</td>
<td>0.4035</td>
</tr>
<tr>
<td>CobaltF-comp</td>
<td>0.0918</td>
<td>0.4691</td>
</tr>
<tr>
<td>DPMFcomb</td>
<td>0.2035</td>
<td>0.4426</td>
</tr>
<tr>
<td>BEER</td>
<td>0.2277</td>
<td>0.3840</td>
</tr>
<tr>
<td>chrF2</td>
<td>0.2177</td>
<td>0.3749</td>
</tr>
</tbody>
</table>

(Work with **Marina Fomicheva**)
Or was it a feature of the uedin systems?

Correlation of various MT systems on 400 sentences per group:

<table>
<thead>
<tr>
<th></th>
<th>PBMT</th>
<th>PBMT + NMT</th>
<th>Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.5662</td>
<td>0.4676</td>
<td>0.4521</td>
</tr>
<tr>
<td>Meteor</td>
<td>0.6178</td>
<td>0.5462</td>
<td>0.5560</td>
</tr>
<tr>
<td>TER</td>
<td>-0.5277</td>
<td>-0.4177</td>
<td>-0.3929</td>
</tr>
<tr>
<td>chrF2</td>
<td>0.5549</td>
<td>0.5093</td>
<td>0.4602</td>
</tr>
<tr>
<td>BEER</td>
<td>0.5445</td>
<td>0.4913</td>
<td>0.4598</td>
</tr>
<tr>
<td>UPF-Cobalt</td>
<td>0.6510</td>
<td><strong>0.5400</strong></td>
<td><strong>0.5221</strong></td>
</tr>
<tr>
<td>CobaltF-comp</td>
<td>0.6328</td>
<td>0.5788</td>
<td>0.5693</td>
</tr>
<tr>
<td>MetricsF</td>
<td>0.6575</td>
<td>0.5840</td>
<td>0.5803</td>
</tr>
<tr>
<td>DPMFcomb</td>
<td>0.6700</td>
<td>0.5876</td>
<td><strong>0.5815</strong></td>
</tr>
</tbody>
</table>

These NMT systems only use neural models for rescoring. Also, average DA scores not higher for the PMT+NMT group.

(Work with Marina Fomicheva)
Conclusions

- (Machine) Translation evaluation is still an open problem
Conclusions

- (Machine) Translation evaluation is still an open problem
- **Quality estimation** and other trained metrics can learn different “versions” of quality
Conclusions

- (Machine) Translation evaluation is still an open problem
- **Quality estimation** and other trained metrics can learn different “versions” of quality
- Which metrics are used in practice?
  - BLEU + your favourite other
  - And same metric for tuning
Conclusions

- (Machine) Translation evaluation is still an open problem
- **Quality estimation** and other trained metrics can learn different “versions” of quality
- Which metrics are used in practice?
  - BLEU + your favourite other
  - And same metric for tuning
- And for **official** comparisons?
  - WMT: manual ranking and direct assessment
  - IWSLT: manual post-editing
Conclusions

- (Machine) Translation evaluation is still an open problem
- **Quality estimation** and other trained metrics can learn different “versions” of quality
- Which metrics are used in practice?
  - BLEU + your favourite other
  - And same metric for tuning
- And for **official** comparisons?
  - WMT: manual ranking and direct assessment
  - IWSLT: manual post-editing
- Are our metrics good at assessing NMT systems?
- Are these metrics good to optimise NMT systems?
Translation Quality Assessment: Evaluation and Estimation

Lucia Specia
University of Sheffield
l.specia@sheffield.ac.uk

MTM - Prague, 12 September 2016
### Conclusions

<table>
<thead>
<tr>
<th>MT system</th>
<th>Type</th>
<th>Average score</th>
<th>Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFRL-MITLL-Phrase</td>
<td>PBMT + NMT</td>
<td>0.0118</td>
<td>56</td>
</tr>
<tr>
<td>AFRL-MITLL-contrast</td>
<td>PBMT + NMT</td>
<td>-0.1423</td>
<td>72</td>
</tr>
<tr>
<td>AMU-UEDIN</td>
<td>PBMT + NMT</td>
<td>0.1981</td>
<td>61</td>
</tr>
<tr>
<td>KIT</td>
<td>PBMT + NMT</td>
<td>0.1431</td>
<td>73</td>
</tr>
<tr>
<td>LIMSI</td>
<td>PBMT</td>
<td>-0.1482</td>
<td>84</td>
</tr>
<tr>
<td>NRC</td>
<td>PBMT</td>
<td>0.0877</td>
<td>58</td>
</tr>
<tr>
<td>PJATK</td>
<td>PBMT</td>
<td>0.0137</td>
<td>132</td>
</tr>
<tr>
<td>PROMT-Rule-based</td>
<td>RBMT</td>
<td>0.0107</td>
<td>56</td>
</tr>
<tr>
<td>PROMT-SMT</td>
<td>PBMT</td>
<td>-0.1163</td>
<td>154</td>
</tr>
<tr>
<td>UH-factored</td>
<td>PBMT</td>
<td>-0.1138</td>
<td>70</td>
</tr>
<tr>
<td>UH-opus</td>
<td>PBMT</td>
<td>-0.0059</td>
<td>72</td>
</tr>
<tr>
<td>cu-mergedtrees</td>
<td>Syntax PBMT</td>
<td>-0.4976</td>
<td>106</td>
</tr>
<tr>
<td>dvorkanton</td>
<td>PBMT + NMT</td>
<td>-0.1548</td>
<td>72</td>
</tr>
<tr>
<td>jhu-pbmt</td>
<td>PBMT</td>
<td>-0.0985</td>
<td>446</td>
</tr>
<tr>
<td>jhu-syntax</td>
<td>Syntax PBMT</td>
<td>-0.2491</td>
<td>125</td>
</tr>
<tr>
<td>online-B</td>
<td>PBMT</td>
<td>0.0793</td>
<td>430</td>
</tr>
<tr>
<td>online-F</td>
<td>PBMT</td>
<td>-0.2447</td>
<td>125</td>
</tr>
<tr>
<td>online-G</td>
<td>PBMT</td>
<td>0.0186</td>
<td>272</td>
</tr>
<tr>
<td>tbtk-syscomb</td>
<td>PBMT</td>
<td>-0.0594</td>
<td>85</td>
</tr>
<tr>
<td>uedin-nmt</td>
<td>NMT</td>
<td>0.0774</td>
<td>342</td>
</tr>
<tr>
<td>uedin-pbmt</td>
<td>PBMT</td>
<td>0.0391</td>
<td>231</td>
</tr>
<tr>
<td>uedin-syntax</td>
<td>Syntax PBMT</td>
<td>0.0121</td>
<td>238</td>
</tr>
</tbody>
</table>