Automated Language Quality Metrics

Gábor Bessenyei (CEO, MorphoLogic Localisation)

✉️ gabor.bessenyei@mloc.eu
Why do we need them?
The growing hunger for content...

... in a global, multilingual economy...

Source: https://artinwords.de/pieter-bruegel-der-aeltere-der-turmbau-zu-babel/
... is resulting in increased demand for translation

Source: https://www.moneycrashers.com/time-banking-trade-services-time-based-currency/
Where do we use them?
Usage scenarios

TM analysis

Post-editing

MT quality

MT QE
The metrics
Different flavors

- Meteor
- Levenshtein
- F-Measure
- PER
- RIBES
- BLEU
- NIST
- TERR
- RED
- CDER
- AMBER

= fuzzy match
In computational linguistics and computer science, edit distance is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other.

The Levenshtein distance between "kitten" and "sitting" is 3. A minimal edit script that transforms the former into the latter is:

1. kitten → sitten (substitution of "s" for "k")
2. sitten → sittin (substitution of "i" for "e")
3. sittin → sitting (insertion of "g" at the end).

BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another.

Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU.

The problem
BLEU (you can replace it with any other metric)

- **BLEU (bilingual evaluation understudy)** is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU. BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics.

- Scores are calculated for individual translated segments—generally sentences—by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. Intelligibility or grammatical correctness are not taken into account.

- BLEU is designed to approximate human judgement at a corpus level, and performs badly if used to evaluate the quality of individual sentences.

- BLEU’s output is always a number between 0 and 1.

Guess the fuzz

80%
I want to be a doctor.

56%

57%
I want to be a translctor.

92%
Guess the fuzz

I want to be a doctor.
I want to be a translator.

- SDL Studio 2015: 89%
- memoQ 2015: 80%
- Memsource: 60%
- Wordfast: 56%
Guess the fuzz

I want to be a doctor.
I want to be a translator.

I want to be a doctor.
I want to be a velociraptor.

SDL Studio 2015 89%
memoQ 2015 80%
Memsorce 60%
Wordfast 56%

SDL Studio 2015 89%
memoQ 2015 80%
Memsorce 55%
Wordfast 53%

* XTM: below 75%, doesn't show
Spot the difference

• One document
• Two CAT tool analysis
What is the match rate?

Please let me go
Me go please let
Let me go please
Which one is better?

**EN original**
Abnormal movement of the levers/linkage indicating maladjustment or excessive wear.

**Reference human translation:**
Übermäßige Hebel-/Gestängewege wegen falscher Einstellung oder übermäßiger Abnutzung

**Globalese NMT:**
Übermäßige Hebel-/Gestängewege wegen falscher Einstellung oder übermäßigen Verschleißes

| 20 | - | 0.4 (4.0/10.0) | Abnormal movement of the levers/linkage indicating maladjustment or excessive wear. | Übermäßige Hebel-/Gestängewege wegen falscher Einstellung oder übermäßiger Abnutzung | Übermäßige Hebel-/Gestängewege wegen falscher Einstellung oder übermäßigen Verschleißes |
Which one is better?

**EN original**
Abnormal movement of the levers/linkage indicating maladjustment or excessive wear.

**Reference human translation:**
Übermäßige Hebel -/ Gestängewege wegen falscher Einstellung oder übermäßiger Abnutzung

**Globalese NMT:**
Übermäßige Hebel -/ Gestängewege wegen falscher Einstellung oder übermäßigen Verschleißes

| 20 | - | 0.4 (4.0/10.0) | Abnormal movement of the levers/linkage indicating maladjustment or excessive wear. | Übermäßige Hebel -/ Gestängewege wegen falscher Einstellung oder übermäßiger Abnutzung | Übermäßige Hebel -/ Gestängewege wegen falscher Einstellung oder übermäßigen Verschleißes |
Which one is better?

Close the machine doors and reset the alarm.

Die Türen schließen und den Alarm quittieren.

Die Maschinentüren schließen und den Alarm quittieren.
<table>
<thead>
<tr>
<th>TER</th>
<th>Source segment</th>
<th>Human translated segment (retokenized)</th>
<th>MT segment (retokenized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.12 (1.0/8.0)</td>
<td>Close the machine doors and reset the alarm.</td>
<td>Die Türen schließen und den Alarm quittieren.</td>
<td>Die Maschinentüren schließen und den Alarm quittieren.</td>
</tr>
</tbody>
</table>
Problem summary

• There is no standard definition of a word or a segment
• There is no standard definition for scaling of matches
• There is no standard how to deal with word order
• Language / translation quality is squeezed into a number between 0 and 1
• A higher number is not necessarily reflecting higher quality
The bottom line

• Automated language quality metrics are far from perfect
• They are still needed and they will be needed
• The language industry should work on standards
Thank you!
Your opinion is important to us! Please tell us what you thought of the lecture. We look forward to your feedback via smartphone or tablet under

http://gala05.honestly.de

or scan the QR code

The feedback tool will be available even after the conference!