GSLT

Machine Translation Evaluation

Eva Forsbom
evafo@stp.ling.uu.se

Uppsala University
Evaluation Standardisation Efforts

ISO
Software
Quality

| EAGLES Framework
| QA Segmentation

/
/
/
/

Writing
Dialogue

Aids
Systems

ISLE
Taxonomy

| SAE
|
|
|

J2450
Quality Attributes

ISO 8402: “The totality of features and characteristics of a product or service that bears on its ability to satisfy stated or implied needs”

ISO/IEC 9126 series: Product quality

ISO/IEC 14598 series: Software product evaluation

- Functionality
- Reliability
- Usability
- Efficiency
- Maintainability
- Portability
Evaluation Context

- For whom?
- Why?
- What?
- By whom?
- How?
For Whom?

Different users have different needs. The quality attributes should be picked and weighted accordingly.

- Consumer agency
- Manager
- Developer
- Experienced user
- Consumer
- ...

GSLT Machine Translation Evaluation – p.5/38
Why?

The purpose of the evaluation depends on the kind of user it is done for, and on the maturity of the product. There is a type of evaluation for each purpose... Some examples:

<table>
<thead>
<tr>
<th>Type</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feasability</td>
<td>See if the product is needed/worth developing</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>Trace errors</td>
</tr>
<tr>
<td>Progressive</td>
<td>See changes between product versions</td>
</tr>
<tr>
<td>Adequacy</td>
<td>See if the product is adequate for a certain task</td>
</tr>
<tr>
<td>Performance</td>
<td>Compare different systems</td>
</tr>
</tbody>
</table>
Depending on user and purpose, attributes at an appropriate level of specificity should be chosen for evaluation. Weighted results for specific attributes could be combined into a higher level attribute.

```
[ functionality : 
  suitability : true
  accuracy : 60%

  interoperability : xx
  security : high
  compliance : true

  reliability : 7
  usability : good
  efficiency : basic
  maintainability : xx
  portability : yy ]
```
By Whom?

The different types of evaluations require different kinds of evaluators with different backgrounds. Some evaluations can be performed automatically, some not.

- Evaluation agency
- Business Manager
- Developer
- Domain expert
- Experienced user
- Bilingual user
- Consumer
- ...

...
How?

The evaluation process can be divided into three general stages:

1. Defining the quality requirements
   - requirements analysis
   - evaluation modelling

2. Preparing the evaluation
   - quality metrics selection
   - rating levels definition
   - assessment criteria definition

3. Proceeding with the evaluation
   - measurement
   - rating
   - assessment
Using ISLE’s MT Evaluation Taxonomy, evaluators can slide down a tree of increasingly specific quality attributes and find appropriate measures for evaluating them. It has two entry points, which are both mapped to the metrics.

1 Specifying user needs
   - The purpose of evaluation
   - The object of evaluation
   - Characteristics of the translation task
     - Assimilation
     - Dissemination
     - Communication
   - User characteristics
   - Input characteristics (author and text)

2 System characteristics to be evaluated
   - System internal characteristics
     - MT system-specific characteristics
     - Model of translation process
     - Linguistic resources and utilities
     - Characteristics of the intended mode
   - System external characteristics
     - Functionality
     - Reliability
     - Usability
     - Efficiency
     - Maintainability
     - Portability
     - Cost
Blackbox Evaluation

In cases where the evaluator has no possibility to see output from the system components, or for high level quality attribute evaluation, a blackbox evaluation is appropriate. Then, only the input, possible settings, and output are known.

<table>
<thead>
<tr>
<th>Input Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
</tr>
<tr>
<td>Total: 11192</td>
</tr>
<tr>
<td>Unique: 2393 (21.38%)</td>
</tr>
<tr>
<td>Segments</td>
</tr>
<tr>
<td>Total: 1772</td>
</tr>
<tr>
<td>Unique: 1187 (66.99%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
</tr>
<tr>
<td>Source Language Words</td>
</tr>
<tr>
<td>Total: 11025 (98.51%)</td>
</tr>
<tr>
<td>Unique: 2322 (97.03%)</td>
</tr>
<tr>
<td>Segments</td>
</tr>
<tr>
<td>Fully Translated</td>
</tr>
<tr>
<td>Total: 594 (33.52%)</td>
</tr>
<tr>
<td>Unique: 210 (17.69%)</td>
</tr>
<tr>
<td>Translated</td>
</tr>
<tr>
<td>Total: 678 (38.26%)</td>
</tr>
<tr>
<td>Unique: 285 (24.01%)</td>
</tr>
</tbody>
</table>
Glassbox Evaluation

In cases where the evaluator has possibility to see output from the system components, or for low level quality attribute evaluation, a glassbox evaluation is appropriate. Then, input, possible settings, and output to some or all components are known.

Error Reports

<table>
<thead>
<tr>
<th></th>
<th>Source Language Words</th>
<th>Translation Links</th>
<th>Target Language Words</th>
<th>Target Language Code</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>167</td>
<td>1795</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td><strong>Unique</strong></td>
<td>71</td>
<td>371</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Not Parsed</th>
<th>Partially Parsed</th>
<th>Not Transferred</th>
<th>Not Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>347</td>
<td>712</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td><strong>Unique</strong></td>
<td>304</td>
<td>577</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>
Translation quality is usually evaluated by comparison of the translated text to the source text (by bilingual evaluators) or to a reference translation (by monolingual evaluators). Some evaluations could be performed automatically.

- Fidelity (how close)
- Correctness (how correct)
- Adequacy (how adequate)
- Informativeness (how informative)
- Intelligibility (how understandable)
- Fluency (how fluent)
Manual Evaluation – Tests

- Grading
- Cloze test
- Comprehension test
- Task-based test
- Reading time
- Typing
- Post-editing
Example: Adequacy Scale

(Doyon, Taylor, and White, 1998)

5 All meaning expressed in the source fragment appears in the translation fragment
4 Most of the source fragment meaning is expressed in the translation fragment
3 Much of the source fragment meaning is expressed in the translation fragment
2 Little of the source fragment meaning is expressed in the translation fragment
1 None of the meaning expressed in the source fragment is expressed in the translation fragment
Example: Adequacy Test for LREC’02

(http://stp.ling.uu.se/~evafo/lrec_eval/)

1 2 3 4 5  Source: Prévenir ses enfants des problèmes de drogue
○ ○ ○ ○ ○  Reference: Prevent your children from having drug problems
Translation: Prevent your children from drug problems
Manual Evaluation – Problems

The hat is fat.
The cat is fat.
The hat is fat.
Semi-automatic evaluation usually involves some form of manual mark-up, followed by automatic comparison and computation, e.g. by certain words, constructions, or information units.

- Named entity translation
- EvalTrans
- Syntactic correctness
- Domain terminology translation
- Information unit translation
- Test suite creation
Example: Named Entity Translation

(Reeder et al. 2001)

In this evaluation, some human annotators marks up named entities (NE) in a reference translation. All unique NE’s from the reference translation are then searched in the translations, and all unique occurrences counted. Some normalisation processes could also be applied.

- Only relevant when many NE’s.
- Depends on the annotators’ consistency.
- Depends on the reference translation quality.
Example: EvalTrans

(Nießen et al. 2000)

EvalTrans is a tool for semi-automatic evaluation of translations. Storing of previous evaluations makes the manual evaluations more consistent.

- Manual seeding of scores (SSER)
- Storing of evaluations (WER and SSER)
- Automatic comparison of new translations with old
- Extrapolation of SSER for new translations
- Highlighting of new translations (with mark-up of edit operations)
- Possibility of splitting segments into information units
Automatic evaluation is usually some form of approximate string matching or a count of mark-ups. If there exist advanced linguistic resources for the languages under scrutiny, much mark-up could be done automatically.

- Edit distance
- N-gram occurrence
- Number of untranslated words
- (Named entity translation)
- (Syntactic correctness)
- (Domain terminology translation)
- (Information unit translation)
- (Test suite creation and evaluation)
Edit Distance – Dynamic Programming

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Edit Distance – Dynamic Programming

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Edit Distance – Dynamic Programming


<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Edit Distance – Dynamic Programming

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Edit Distance – Dynamic Programming

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
## Edit Distance – Aligning

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>*</th>
<th>*</th>
<th>*</th>
<th>*</th>
<th>*</th>
<th>*</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>*</td>
<td>*</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>
<td>*</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>i</td>
</tr>
</tbody>
</table>
## Edit Distance – Aligning

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

```
* * * * * R S
* * * * * * *
* * * * * * *
* * * * * i i
```
## Edit Distance – Aligning

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

```
* * * * E R S
* * * * E *
* * * * c i i
```
**Edit Distance – Aligning**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccccc}
* & * & * & D & E & R & S \\
* & * & * & E & * & * \\
* & * & * & i & c & i & i \\
\end{array}
\]
# Edit Distance – Aligning

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* * N D E R S
* * N * E * *
* * c i c i i
### Edit Distance – Aligning

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>*</th>
<th>*</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>R</td>
<td>N</td>
<td>*</td>
<td>E</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>d</td>
<td>c</td>
<td>i</td>
<td>c</td>
<td>i</td>
<td>i</td>
<td></td>
</tr>
</tbody>
</table>
## Edit Distance – Aligning

<table>
<thead>
<tr>
<th>A</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>*</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>R</td>
<td>N</td>
<td>*</td>
<td>E</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>c</td>
<td>d</td>
<td>c</td>
<td>i</td>
<td>c</td>
<td>i</td>
<td>i</td>
</tr>
</tbody>
</table>
Example: Word Accuracy

(Alshawi et al. 1998)

\[
WA = \left( 1 - \frac{d + s + i}{r} \right)
\]

where

\[
\begin{align*}
    d &= \text{deletions} \\
    s &= \text{substitutions} \\
    i &= \text{insertions} \\
    r &= \text{length of reference}
\end{align*}
\]
The original word accuracy measure could result in a score less than 0, as in the following example:

**Src:** Tätningsring  
**Cand:** Sealing ring  
**Ref:** Seal

\[
\left( 1 - \frac{1 + 1 + 0}{1} \right) = -1
\]
Revised Word Accuracy

\[
\text{WA}_{\text{rev}} = \left(1 - \frac{d + s + i}{\max(r, c)}\right)
\]

where

\[
\begin{align*}
d &= \text{deletions} \\
s &= \text{substitutions} \\
i &= \text{insertions} \\
r &= \text{length of reference} \\
c &= \text{length of candidate}
\end{align*}
\]
Word Accuracy vs. Revised Word Accuracy

- BLEU/N(min)–MEAN
- WA/WArev
- BLEU/WA
- N(min)–MEAN/WArev

GSLT Machine Translation Evaluation – p.27/38
Word Accuracy Weaknesses

- Sensitive to word order reversal
- Only evaluated against one reference translation at a time

**Src:** Cylinder, underdel  
**Cand:** Bottom cylinder  
**Ref:** Cylinder bottom

**Src:** Ledningsnät för bränslepump  
**Cand:** Cable harness for fuel pump  
**Ref:** Fuel pump cable harness
N-Gram Occurrence

N-gram occurrence is a way of measuring if words are correctly translated (1-grams) and if the translation is idiomatic \((n > 1)\). It seems to correlate well with human evaluation of accuracy and fluency.

**BLEU (Papineni et al. 2001)**

- Grade = \([0, 1]\);
- Compensates for difference in length by a brevity penalty;
- Applies equal weights for all n-grams.

**NIST (DARPA 2001(?))**

- Grade = \([0, \infty)\);
- Compensates for difference in length by another brevity penalty;
- Applies different weights for the n-grams.
Example: BLEU

\[
\text{BLEU} = BP \cdot \exp \left( \sum_{n=1}^{N} w_n \log p_n \right)
\]

where

\[
BP = \begin{cases} 
1 & \text{if } c > r \\
\exp \left( 1 - \frac{r}{c} \right) & \text{if } c \leq r 
\end{cases}
\]

\(r = \text{length of reference}\)

\(c = \text{length of candidate}\)

\(N = 4\)

\(w = \frac{1}{N}\)

\[p = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n \in \{\text{Candidates}\}} \text{Count}_{\text{clip}}(n)}{\sum_{C \in \{\text{Candidates}\}} \sum_{n \in \{\text{Candidates}\}} \text{Count}(n)}\]
The original BLEU measure is not defined for all cases, as in the following examples:

**Src:** Cylinder, underdel  
**Cand:** Bottom cylinder  
**Ref:** Cylinder bottom

**Src:** Ledningsnät för bränslepump  
**Cand:** Cable harness for fuel pump  
**Ref:** Fuel pump cable harness
N-MEAN – Revised BLEU

\[
N \text{-MEAN} = BP \cdot \sum_{n=1}^{N} w_n p_n
\]

where

\[
N = \begin{cases} 
N_{\text{max}} & \text{if } c \geq N_{\text{max}} \\
 c & \text{if } c < N_{\text{max}} 
\end{cases}
\]
N-Gram Occurrence Weakness

- Sensitive to word errors (particularly mid-segment)

**Cand:** The cats is fat

**Ref:** The cat is fat
Ongoing and Future Work

- Applying these automatic measures on another text type
- Applying these automatic measures on another domain
- Applying these automatic measures on another language pair
- Applying these automatic measures with only one reference translation
- Using other automatic measures
- Using more linguistic measures
References


- DARPA. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, 2001(?).


- EAGLES (Expert Advisory Group on Language Engineering Standards)
  
- ISLE (International Standards for Language Engineering)
  http://www.issco.unige.ch/projects/isle

- ISO (International Organization for Standardization)
  http://www.iso.org

- LISA (Localization Industry Standards Association)
  http://www.lisa.org

References...


- SAE (Society of Automotive Engineers).
  
  http://www.sae.org/