Adequacy-Fluency Metrics (AM-FM) for Machine Translation (MT) Evaluation

Haizhou Li

Invited Talk: WAT 2015, 16 Oct 2015, Kyoto, Japan

Agenda

- The evaluation of ASR and MT
- How do machines evaluate translations today?
- How do humans evaluate translations?
- The Adequacy-Fluency Metrics (AM-FM)
- The mathematical formulation
- The experiments
Automatic Evaluation of Automatic Speech Recognition

ASR output is compared to a reference transcription.

The reference transcription is unique!
Automatic Evaluation of Machine Translation

MT output is compared to reference translations.

... but references are not unique!

文无第一，武无第二
Agenda

• The evaluation of ASR and MT
• How do machines evaluate translations today?
• How do humans evaluate translations?
• The Adequacy-Fluency Metrics (AM-FM)
• The mathematical formulation
• The experiments
Traditional Evaluation Approach

Compare the output with a set of references

WER\(^1\), PER\(^1\) → Compare words

BLEU\(^2\), NIST\(^3\) → Compare n-grams


**Traditional Approach: Good Scores**

Translation Output: This is a toilet.

Reference Translation: This is a toilet.

- **word matches = 4/4**
- **n-gram matches = 5/5**

**Good Score** → **Good Translation**
**Traditional Approach: Bad Scores**

Translation Output ➔ It’s the Water Closet.

Reference Translation ➔ This is a toilet.

word matches = 0/4  
n-gram matches = 0/5  
Bad Score ➔ ?
Traditional Approach: Better Scores?

Translation Output ➔ This isn’t a toilet.

Reference Translation ➔ This is a toilet.

word matches = 3/4

\( n \)-gram matches = 3/5

Better Score ➔ ?
How Machines Evaluate Translations?

- Only look at outputs and references
- Without knowledge support
A Semantic Framework is Needed

Automatic MT evaluation must move beyond words and $n$-grams! Some recent proposals:

- **METEOR**\(^1\)
  - Compare stems and synonyms
- **TER**\(^2\)
  - Compute edit distances
- **MEANT**\(^3\)
  - Compare semantic roles


2.- M. Snover et al., “Study of Translation Edit Rate with Targeted Human Annotation”, in *Proc. of the 7th Biennial Conf. of the Assoc. for Mach. Translation in the Amer.*, Cambridge, MA, USA, Aug 2006

Agenda

• The evaluation of ASR and MT
• How do machines evaluation translations today?
• How do humans evaluation translations?
• The Adequacy-Fluency Metrics (AM-FM)
• The mathematical formulation
• The experiments
How Humans Evaluate Translations?* (I)

ADEQUACY
How much of the source information is preserved?

P(T|S) ≈ P(S|T) P(T)

FLUENCY
How good is the generated target language quality?

How Humans Evaluate Translations? (II)

• Look at both outputs and inputs

• Language and cultural knowledge
Adequacy Evaluation Scale*

How much of the source information is preserved in the translation?
(Look at both inputs and outputs!)

<table>
<thead>
<tr>
<th>Score</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>None of the meaning is preserved</td>
</tr>
<tr>
<td>2</td>
<td>Little of the meaning is preserved</td>
</tr>
<tr>
<td>3</td>
<td>Much of the meaning is preserved</td>
</tr>
<tr>
<td>4</td>
<td>Most of the meaning is preserved</td>
</tr>
<tr>
<td>5</td>
<td>All the meaning is preserved</td>
</tr>
</tbody>
</table>

**Fluency Evaluation Scale**

How good is translation regarding the target language quality?

(Only look at the outputs!)

<table>
<thead>
<tr>
<th>Score</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Incomprehensible target language</td>
</tr>
<tr>
<td>2</td>
<td>Disfluent target language</td>
</tr>
<tr>
<td>3</td>
<td>Non-native kind of target language</td>
</tr>
<tr>
<td>4</td>
<td>Good quality target language</td>
</tr>
<tr>
<td>5</td>
<td>Flawless target language</td>
</tr>
</tbody>
</table>

Agenda

• The evaluation of ASR and MT
• How do machines evaluate translations today?
• How do humans evaluate translations?
• The Adequacy-Fluency Metrics (AM-FM)
• The mathematical formulation
• The experiments
The Proposed Evaluation Framework*

• Approximate adequacy and fluency by means of independent models:
  – Use a “semantic approach” for adequacy
  – Use a “syntactic approach” for fluency

• Combine both evaluation metrics into a single evaluation score

# State of the Art in MT Evaluation*

<table>
<thead>
<tr>
<th>Assessment Level</th>
<th>Need for References</th>
<th>Cross-Language Approach</th>
<th>Humans in the Loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>WER, PER</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Word $n$-grams</td>
<td>BLEU, NIST</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stems &amp; Synonyms</td>
<td>METEOR</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Edit Distances</td>
<td>TER</td>
<td>-</td>
<td>HTER</td>
</tr>
<tr>
<td>Semantic Roles</td>
<td>MEANT</td>
<td>XMEANT</td>
<td>HMEANT</td>
</tr>
<tr>
<td>Continuous Space</td>
<td>$m$AM-FM</td>
<td>$x$AM-FM</td>
<td>-</td>
</tr>
</tbody>
</table>

Properties of Continuous Spaces

The Distributional Hypothesis

“a word is characterized for the company it keeps” (Firth 1957)
meaning is mainly determined by the context rather than from individual language units

- Continuous spaces represent semantic similarities by means of the geometric concept of proximity
- Offer much “better” smoothing capabilities
- Not constrained to the Markovian assumption
The Term-Document Matrix

- A model representing joint distributions between words and documents

<table>
<thead>
<tr>
<th></th>
<th>D_1</th>
<th>D_2</th>
<th>D_3</th>
<th>D_4</th>
<th>D_5</th>
<th>D_6</th>
<th>D_7</th>
<th>D_8</th>
<th>...</th>
<th>D_N</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T_3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T_4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T_5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T_6</td>
<td></td>
<td></td>
<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>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T_M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Non-zero row values for those documents containing a given word

Non-zero column values for those words occurring within a given document
**Document Vector Spaces**

Pay attention to the columns of the term-document matrix.

<table>
<thead>
<tr>
<th>Variables</th>
<th>D₁</th>
<th>D₂</th>
<th>D₃</th>
<th>D₄</th>
<th>D₅</th>
<th>D₆</th>
<th>D₇</th>
<th>D₈</th>
<th>...</th>
<th>Dₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T₃</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T₄</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T₅</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T₆</td>
<td></td>
<td></td>
<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>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tₘ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Semantic Association in Vector Spaces

Association scores and similarity metrics can be used to assess the degree of semantic relatedness among documents.
Semantic Map for Data Collection (1)

Opinionated content from rating website

Positive

Negative

Automotive

Financial
Semantic Map for Data Collection (2)

66 Books from The Holy Bible: English version

(vocabulary size: 8121 words)
Agenda

• The evaluation of ASR and MT
• How do machines evaluate translations today?
• How do humans evaluate translations?
• The Adequacy-Fluency Metrics (AM-FM)
• The mathematical formulation
• The experiments
**AM: Adequacy-oriented Metric**

- Compare sentences in a semantic space
  - Monolingual AM ($mAM$): compare output vs. reference
  - Cross-language AM ($xAM$): compare output vs. input
Latent Semantic Indexing (LSI)*

SVD: \( \mathbf{M}_{M \times N} = \mathbf{U}_{M \times M} \Sigma_{M \times N} \mathbf{V}^T_{N \times N} \)

Documents projected into word space

\( \mathbf{U}^T_{M \times M} \mathbf{M}_{M \times N} = \mathbf{D}_{M \times M} \)

Documents projected into reduced word (semantic) space

Translation output (\( \text{to} \)) and translation reference (\( \text{tr} \)) compared in reduced vector space

\( \langle \mathbf{U}^T_{K \times M} \text{to}_{M \times 1}, \mathbf{U}^T_{K \times M} \text{tr}_{M \times 1} \rangle \)

Cross-Language LSI*

\[ X_{(Ms+Mt)\times N} = \begin{pmatrix} M_{Ms \times N} \\ M_{Mt \times N} \end{pmatrix} \]

**SVD:** \[ X = U \Sigma V^T \]

Translation output (to) and translation input (ti) compared in cross-language vector space

\[ <U^T_{K \times (Ms+Mt)} \begin{bmatrix} 0_{Ms \times 1} \\ to_{Mt \times 1} \end{bmatrix}, U^T_{K \times (Ms+Mt)} \begin{bmatrix} ti_{Ms \times 1} \\ 0_{Mt \times 1} \end{bmatrix}> \]

**FM: Fluency-oriented Metric**

- Measures the quality of the target language with a language model.
- Uses a compensation factor to avoid effects derived from differences in sentence lengths.
Compensated Language Model

\[ FM = \exp \left( \frac{1}{N} \sum_{n=1:N} \log( p(w_n | w_{n-1}, \ldots)) \right) \]

*n*-gram probabilities

compensation factor
**AM-FM Combined Score**

Both components can be combined into a single metric according to different criteria

- **Weighted Harmonic Mean:**
  \[
  H-AM-FM = \frac{AM \cdot FM}{\alpha AM + (1-\alpha) FM}
  \]

- **Weighted Mean:**
  \[
  M-AM-FM = (1-\alpha) AM + \alpha FM
  \]

- **Weighted L2-norm:**
  \[
  N-AM-FM = \sqrt{(1-\alpha) AM^2 + \alpha FM^2}
  \]
Agenda

• The evaluation of ASR and MT
• How do machines evaluate translations today?
• How do humans evaluate translations?
• The Adequacy-Fluency Metrics (AM-FM)
• The mathematical formulation
• The experiments
**WMT-2007 Dataset**

- Fourteen tasks:
  - five European languages (EN, ES, DE, FR, CZ) and
  - two different domains (News and EPPS).
- Systems outputs available from 14 teams that had participated in the evaluation. In total, 86 system outputs.
- Overall 172,315 individual sentence translations, from which a total of 10,754 were rated for both adequacy and fluency by human judges.

## WMT-2007 Translation Task Details

<table>
<thead>
<tr>
<th>Task</th>
<th>Domain</th>
<th>Source</th>
<th>Target</th>
<th>Systems</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>News</td>
<td>CZ</td>
<td>EN</td>
<td>3</td>
<td>727</td>
</tr>
<tr>
<td>T2</td>
<td>News</td>
<td>EN</td>
<td>CZ</td>
<td>2</td>
<td>806</td>
</tr>
<tr>
<td>T3</td>
<td>EPPS</td>
<td>EN</td>
<td>FR</td>
<td>7</td>
<td>577</td>
</tr>
<tr>
<td>T4</td>
<td>News</td>
<td>EN</td>
<td>FR</td>
<td>8</td>
<td>561</td>
</tr>
<tr>
<td>T5</td>
<td>EPPS</td>
<td>EN</td>
<td>DE</td>
<td>6</td>
<td>924</td>
</tr>
<tr>
<td>T6</td>
<td>News</td>
<td>EN</td>
<td>DE</td>
<td>6</td>
<td>892</td>
</tr>
<tr>
<td>T7</td>
<td>EPPS</td>
<td>EN</td>
<td>ES</td>
<td>6</td>
<td>703</td>
</tr>
<tr>
<td>T8</td>
<td>News</td>
<td>EN</td>
<td>ES</td>
<td>7</td>
<td>832</td>
</tr>
<tr>
<td>T9</td>
<td>EPPS</td>
<td>FR</td>
<td>EN</td>
<td>7</td>
<td>624</td>
</tr>
<tr>
<td>T10</td>
<td>News</td>
<td>FR</td>
<td>EN</td>
<td>7</td>
<td>740</td>
</tr>
<tr>
<td>T11</td>
<td>EPPS</td>
<td>DE</td>
<td>EN</td>
<td>7</td>
<td>949</td>
</tr>
<tr>
<td>T12</td>
<td>News</td>
<td>DE</td>
<td>EN</td>
<td>5</td>
<td>939</td>
</tr>
<tr>
<td>T13</td>
<td>EPPS</td>
<td>ES</td>
<td>EN</td>
<td>8</td>
<td>812</td>
</tr>
<tr>
<td>T14</td>
<td>News</td>
<td>ES</td>
<td>EN</td>
<td>7</td>
<td>668</td>
</tr>
</tbody>
</table>
Metric Correlation with Human Scores

Pearson’s correlation coefficients between the $mAM-FM$ Weighted Mean (left) and $xAM-FM$ Weighted Mean (right) components and human-generated scores for adequacy.
mAM-FM and Adequacy
mAM-FM and Fluency
xAM-FM and Adequacy
xAM-FM and Fluency
### Comparative Evaluation Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>$\alpha$</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>-</td>
<td>0.4107</td>
<td>0.4432</td>
</tr>
<tr>
<td>Meteor</td>
<td>-</td>
<td>0.3505</td>
<td>0.3626</td>
</tr>
<tr>
<td>NIST</td>
<td>-</td>
<td>0.3226</td>
<td>0.3444</td>
</tr>
<tr>
<td>TER-Plus</td>
<td>-</td>
<td>0.3068</td>
<td>0.3170</td>
</tr>
<tr>
<td>$m$AM</td>
<td>-</td>
<td>0.3435</td>
<td>0.3245</td>
</tr>
<tr>
<td>$x$AM</td>
<td>-</td>
<td>0.1291*</td>
<td>0.0330*</td>
</tr>
<tr>
<td>FM</td>
<td>-</td>
<td>0.3408</td>
<td>0.4267</td>
</tr>
<tr>
<td>$m$AM-FM$_{HM}$</td>
<td>0.10</td>
<td>0.4473</td>
<td>0.4977</td>
</tr>
<tr>
<td>$m$AM-FM$_{WM}$</td>
<td>0.60</td>
<td>0.4574</td>
<td>0.5036</td>
</tr>
<tr>
<td>$m$AM-FM$_{L2}$</td>
<td>0.86</td>
<td>0.4523</td>
<td>0.5040</td>
</tr>
<tr>
<td>$x$AM-FM$_{HM}$</td>
<td>0.30</td>
<td>0.4091</td>
<td>0.4503</td>
</tr>
<tr>
<td>$x$AM-FM$_{WM}$</td>
<td>0.60</td>
<td>0.4167</td>
<td>0.4442</td>
</tr>
<tr>
<td>$x$AM-FM$_{L2}$</td>
<td>0.80</td>
<td>0.4084</td>
<td>0.4493</td>
</tr>
</tbody>
</table>

All coefficients (except those marked with ‘*’) are significant with $p<0.01$.

**Updated:** 18/10/2015
Human Adequacy and Fluency

\[ \rho = 0.699 \]
AM and FM Metrics

\[ \rho = 0.035 \]
Conclusions

• We have proposed a new evaluation framework for MT evaluation operating on a continuous space

• mAM-FM achieve better correlations with human evaluations for both adequacy and fluency than other conventional metrics

• xAM-FM allows for quality assessment without the need for a set of reference translations, its performance is still comparable to other state-of-the-art automatic evaluation metrics
Thank You

Online:
www.i2r.a-star.edu.sg
www.facebook.com/i2r.research