Machine Translation Evaluation

Sara Stymne

2017-03-29

Partly based on Philipp Koehn’s slides for chapter 8
Why Evaluation?

- How good is a given machine translation system?
- Which one is the best system for our purpose?
- How much did we improve our system?
- How can we tune our system to become better?
- Hard problem, since many different translations acceptable → semantic equivalence / similarity
Ten Translations of a Chinese Sentence

这个 机场 的 安全 工作 由 以色列 方面 负责。

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport’s security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport’s security is the responsibility of the Israeli security officials.

(a typical example from the 2001 NIST evaluation set)
Which translation is best?

Source  Färjetransporterna har minskat med 20,3 procent i år.
Gloss  The-ferry-transports have decreased by 20.3 percent in year.
Ref    Ferry transports are down by 20.3% in 2008.
Which translation is best?

**Source**  Färjetransporterna har minskat med 20,3 procent i år.

**Gloss**  The-ferry-transports have decreased by 20.3 percent in year.

**Ref**  Ferry transports are down by 20.3% in 2008.

**Sys1**  The ferry transports has reduced by 20.3% in year.

**Sys2**  This year, the reduction of transports by ferry is 20,3 procent.

**Sys3**  Färjetransporterna are down by 20.3% this year.

**Sys4**  Ferry transports have a reduction of 20.3 percent in year.

**Sys5**  Transports are down by 20.3%.
Evaluation Methods

- Subjective judgments by human evaluators
- Task-based evaluation, e.g.:
  - How much post-editing effort?
  - Does information come across?
- Automatic evaluation metrics
- Quality estimation
Human vs Automatic Evaluation

- **Human evaluation is**
  - Ultimately what we are interested in, but
  - Very time consuming
  - Not re-usable
  - Subjective

- **Automatic evaluation is**
  - Cheap and re-usable, but
  - Not necessarily reliable
Human evaluation

- Adequacy/Fluency (1 to 5 scale)
- Ranking of systems (best to worst)
- Yes/no assessments (acceptable translation?)
- SSER – subjective sentence error rate ("perfect" to "absolutely wrong")
- Usability (Good, useful, useless)
- Human post-editing time
- Error analysis
Adequacy and Fluency

- given: machine translation output
- given: source and/or reference translation
- task: assess the quality of the machine translation output

**Adequacy:** Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?

**Fluency:** Is the output good fluent target language? This involves both grammatical correctness and idiomatic word choices.
### Fluency and Adequacy: Scales

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>all meaning</td>
</tr>
<tr>
<td>4</td>
<td>flawless English</td>
</tr>
<tr>
<td>3</td>
<td>most meaning</td>
</tr>
<tr>
<td>4</td>
<td>good English</td>
</tr>
<tr>
<td>3</td>
<td>much meaning</td>
</tr>
<tr>
<td>3</td>
<td>non-native English</td>
</tr>
<tr>
<td>2</td>
<td>little meaning</td>
</tr>
<tr>
<td>2</td>
<td>disfluent English</td>
</tr>
<tr>
<td>1</td>
<td>none</td>
</tr>
<tr>
<td>1</td>
<td>incomprehensible</td>
</tr>
</tbody>
</table>
Färjetransporterna har minskat med 20,3 procent i år.

The-ferry-transports have decreased by 20.3 percent in year.

Ferry transports are down by 20.3% in 2008.

Ferry transports have a reduction of 20.3 percent in year.

Transports are down by 20.3%.

This year, of transports by ferry reduction is percent 20.3.
Evaluators Disagree

- Histogram of adequacy judgments by different human evaluators

(from WMT 2006 evaluation)
Measuring Agreement between Evaluators

- **Kappa coefficient**

\[
K = \frac{p(A) - p(E)}{1 - p(E)}
\]

- \( p(A) \): proportion of times that the evaluators agree
- \( p(E) \): proportion of time that they would agree by chance

- **Example:** Inter-evaluator agreement in WMT 2007 evaluation campaign

<table>
<thead>
<tr>
<th>Evaluation type</th>
<th>( P(A) )</th>
<th>( P(E) )</th>
<th>( K )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>.400</td>
<td>.2</td>
<td>.250</td>
</tr>
<tr>
<td>Adequacy</td>
<td>.380</td>
<td>.2</td>
<td>.226</td>
</tr>
</tbody>
</table>
Task for evaluator: Is translation X better than translation Y?
(choices: better, worse, equal)

Evaluators are more consistent:

<table>
<thead>
<tr>
<th>Evaluation type</th>
<th>(P(A))</th>
<th>(P(E))</th>
<th>(K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>.400</td>
<td>.2</td>
<td>.250</td>
</tr>
<tr>
<td>Adequacy</td>
<td>.380</td>
<td>.2</td>
<td>.226</td>
</tr>
<tr>
<td>Sentence ranking</td>
<td>.582</td>
<td>.333</td>
<td>.373</td>
</tr>
</tbody>
</table>
Error Analysis

- Analysis and classification of the errors from an MT system
- Many general frameworks for classification exists, e.g.
  - Flanagan, 1994
  - Vilar et al. 2006
  - Costa-jussà et al. 2012
- It is also possible to analyse specific phenomena, like compound translation, agreement, pronoun translation, ...
Example Error Typology

Vilar et al.
Task-Oriented Evaluation

- Machine translations is a means to an end
- Does machine translation output help accomplish a task?
- Example tasks
  - producing high-quality translations post-editing machine translation
  - information gathering from foreign language sources
Post-Editing Machine Translation

- Measuring time spent on producing translations
  - baseline: translation from scratch
  - post-editing machine translation
- Some issues:
  - time consuming
  - depends on skills of translator/post-editor
Content Understanding Tests

Given machine translation output, can monolingual target side speaker answer questions about it?

1. basic facts: who? where? when? names, numbers, and dates
2. actors and events: relationships, temporal and causal order
3. nuance and author intent: emphasis and subtext

Very hard to devise questions
Goals for Evaluation Metrics

**Low cost:** reduce time and money spent on carrying out evaluation

**Tunable:** automatically optimize system performance towards metric

**Meaningful:** score should give intuitive interpretation of translation quality

**Consistent:** repeated use of metric should give same results

**Correct:** metric must rank better systems higher
Other Evaluation Criteria

When deploying systems, considerations go beyond quality of translations

**Speed:** we prefer faster machine translation systems

**Size:** fits into memory of available machines (e.g., handheld devices)

**Integration:** can be integrated into existing workflow

**Customization:** can be adapted to user’s needs
Automatic Evaluation Metrics

- Goal: computer program that computes the quality of translations
- Advantages: low cost, tunable, consistent

- Basic strategy
  - given: machine translation output
  - given: human reference translation
  - task: compute similarity between them
Metrics – overview

- Precision-based
  - BLEU, NIST, …

- F-score-based
  - Meteor, …

- Error rates
  - WER, TER, PER, …

- Using syntax/semantics
  - PosBleu, Meant, DepRef, …

- Using machine learning
  - SVM-based techniques, TerrorCat
Metrics – overview

- Precision-based
  - **BLEU, NIST, …**
- F-score-based
  - **Meteor, …**
- Error rates
  - **WER, TER, PER, …**
- Using syntax/semantics
  - PosBleu, Meant, DepRef, …
- Using machine learning
  - SVM-based techniques, TerrorCat
Precision and Recall of Words

SYSTEM A: Israeli officials responsibility of airport safety

REFERENCE: Israeli officials are responsible for airport security

- **Precision**
  \[
  \frac{\text{correct}}{\text{output-length}} = \frac{3}{6} = 50\%
  \]

- **Recall**
  \[
  \frac{\text{correct}}{\text{reference-length}} = \frac{3}{7} = 43\%
  \]

- **F-measure**
  \[
  \frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%
  \]
Precision and Recall

SYSTEM A: Israeli officials responsibility of airport safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

<table>
<thead>
<tr>
<th>Metric</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>recall</td>
<td>43%</td>
<td>86%</td>
</tr>
<tr>
<td>f-measure</td>
<td>46%</td>
<td>92%</td>
</tr>
</tbody>
</table>

flaw: no penalty for reordering
BLEU

- N-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

$$\text{BLEU} = \min \left( 1, \frac{\text{output-length}}{\text{reference-length}} \right) \left( \prod_{i=1}^{4} \text{precision}_i \right)^{\frac{1}{4}}$$

- Typically computed over the entire corpus, not single sentences
Example

<table>
<thead>
<tr>
<th>Metric</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision (1gram)</td>
<td>3/6</td>
<td>6/6</td>
</tr>
<tr>
<td>precision (2gram)</td>
<td>1/5</td>
<td>4/5</td>
</tr>
<tr>
<td>precision (3gram)</td>
<td>0/4</td>
<td>2/4</td>
</tr>
<tr>
<td>precision (4gram)</td>
<td>0/3</td>
<td>1/3</td>
</tr>
<tr>
<td>brevity penalty</td>
<td>6/7</td>
<td>6/7</td>
</tr>
<tr>
<td>BLEU</td>
<td>0%</td>
<td>52%</td>
</tr>
</tbody>
</table>
Multiple Reference Translations

- To account for variability, use multiple reference translations
  - n-grams may match in any of the references
  - closest reference length used (usually)

- Example

  SYSTEM:
  
<table>
<thead>
<tr>
<th>Israeli officials</th>
<th>responsibility of</th>
<th>airport</th>
<th>safety</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-GRAM MATCH</td>
<td>2-GRAM MATCH</td>
<td>1-GRAM</td>
<td></td>
</tr>
</tbody>
</table>

  Israeli officials are responsible for airport security
  Israel is in charge of the security at this airport
  The security work for this airport is the responsibility of the Israel government
  Israeli side was in charge of the security of this airport
NIST

- Similar to Bleu in that it measures N-gram precision
- Differences:
  - Arithmetic mean (not geometric)
  - Less frequent n-grams are weighted more heavily
  - Different brevity penalty
  - $N = 5$
Partial credit for matching stems

SYSTEM  Jim walk home
REFERENCE  Joe walks home

Partial credit for matching synonyms

SYSTEM  Jim strolls home
REFERENCE  Joe walks home

Use of paraphrases

Different weights for content and function words (later versions)
- Both recall and precision
- Only unigrams (not higher n-grams)
- Flexible matching (Weighted P and R)
- Fluency captured by a penalty for high number of chunks

\[ F_{\text{mean}} = \frac{PR}{\alpha \cdot P + (1 - \alpha) \cdot R} \]

\[ Penalty = 0.5 \cdot \gamma \cdot \left( \frac{\#\text{chunks}}{\#\text{unigrams\_matched}} \right)^\beta \]

\[ \text{Meteor} = (1 - \text{Penalty}) \cdot F_{\text{mean}} \]
METEOR: tuning

- Meteor parameters can be tuned based on human judgments

<table>
<thead>
<tr>
<th>Language</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$\delta$</th>
<th>$w_{\text{exact}}$</th>
<th>$w_{\text{stem}}$</th>
<th>$w_{\text{syn}}$</th>
<th>$w_{\text{par}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universal</td>
<td>.70</td>
<td>1.40</td>
<td>.30</td>
<td>.70</td>
<td>1.00</td>
<td>–</td>
<td>–</td>
<td>.60</td>
</tr>
<tr>
<td>English</td>
<td>.85</td>
<td>.20</td>
<td>.60</td>
<td>.75</td>
<td>1.00</td>
<td>.60</td>
<td>.80</td>
<td>.60</td>
</tr>
<tr>
<td>French</td>
<td>.90</td>
<td>1.40</td>
<td>.60</td>
<td>.65</td>
<td>1.00</td>
<td>.20</td>
<td>–</td>
<td>.40</td>
</tr>
<tr>
<td>German</td>
<td>.95</td>
<td>1.00</td>
<td>.55</td>
<td>.55</td>
<td>1.00</td>
<td>.80</td>
<td>–</td>
<td>.20</td>
</tr>
</tbody>
</table>
Word Error Rate

- Minimum number of editing steps to transform output to reference
  - match: words match, no cost
  - substitution: replace one word with another
  - insertion: add word
  - deletion: drop word
- Levenshtein distance

\[
\text{WER} = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{reference-length}}
\]
Example

<table>
<thead>
<tr>
<th>Metric</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>word error rate (WER)</td>
<td>57%</td>
<td>71%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>airport security</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Israeli officials</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>are</td>
<td>2 2 2 3 4</td>
<td>2 2 2 3 4</td>
</tr>
<tr>
<td>responsible</td>
<td>3 3 3 3 4</td>
<td>2 3 2 3 4</td>
</tr>
<tr>
<td>for</td>
<td>4 4 4 4 3</td>
<td>4 3 4 3 2</td>
</tr>
<tr>
<td>airport security</td>
<td>5 5 5 5 4</td>
<td>5 4 5 4 3</td>
</tr>
<tr>
<td>Israeli officials</td>
<td>6 6 6 6 5</td>
<td>6 6 6 5 4</td>
</tr>
<tr>
<td>are</td>
<td>7 7 7 7 5</td>
<td>7 7 7 7 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>airport security</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Israeli officials</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>are</td>
<td>2 2 2 3 4</td>
<td>2 2 2 3 4</td>
</tr>
<tr>
<td>responsible</td>
<td>3 3 3 3 4</td>
<td>2 3 2 3 4</td>
</tr>
<tr>
<td>for</td>
<td>4 4 4 4 3</td>
<td>4 3 4 3 2</td>
</tr>
<tr>
<td>airport security</td>
<td>5 5 5 5 4</td>
<td>5 4 5 4 3</td>
</tr>
<tr>
<td>Israeli officials</td>
<td>6 6 6 6 5</td>
<td>6 6 6 5 4</td>
</tr>
<tr>
<td>are</td>
<td>7 7 7 7 5</td>
<td>7 7 7 7 5</td>
</tr>
</tbody>
</table>
Other error rates

- **PER** – position-independent word error rate
  - Does not consider the order of words

- **TER** – translation edit rate
  - Adds the operation SHIFT – the movement of a contiguous sequence of words an arbitrary distance

- **SER** – sentence error rate
  - The percentage of sentences that are identical to reference sentences
Metrics using syntax/semantics

- Posbleu, Bleu calculated on part-of-speech
- ULC – Overlap of:
  - shallow parsing
  - dependency and constituent parsing
  - named entities
  - semantic roles
  - discourse representation structures
- Using dependency structures
- Meant, semantic roles
- Considerations:
  - parsers/taggers do not perform well on misformed MT output
  - parsers/tagger not available for all languages
Critique of Automatic Metrics

- Ignore relevance of words
  (names and core concepts more important than determiners and punctuation)
- Operate on local level
  (do not consider overall grammaticality of the sentence or sentence meaning)
- Scores are meaningless
  (scores very test-set specific, absolute value not informative)
- Human translators score low on BLEU
  (possibly because of higher variability, different word choices)
Evaluation of Evaluation Metrics

- Automatic metrics are low cost, tunable, consistent
- But are they correct?
  → Yes, if they correlate with human judgement
Correlation with Human Judgement

![Correlation Graph]

- **Adequacy** (Red Diamonds)
- **Fluency** (Green Circles)

Regression Lines:
- Adequacy: $R^2 = 88.0\%$
- Fluency: $R^2 = 90.2\%$
Evidence of Shortcomings of Automatic Metrics

Post-edited output vs. statistical systems (NIST 2005)

![Graph showing human score vs. Bleu score with correlation line]
Evidence of Shortcomings of Automatic Metrics

Rule-based vs. statistical systems

- Human Score vs. Bleu Score for SMT System 1 and SMT System 2.
- Rule-based System (Systran) compared.

Adequacy vs. Fluency

Graph showing performance metrics of different systems.
Metric Research

- Active development of new metrics
  - syntactic similarity
  - semantic equivalence or entailment
  - metrics targeted at reordering
  - trainable metrics
  - etc.

- Evaluation campaigns that rank metrics
  (using Pearson’s correlation coefficient)
Correlations of metrics with human ranking

<table>
<thead>
<tr>
<th>Metric</th>
<th>de-en</th>
<th>en-de</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>.90</td>
<td>.79</td>
</tr>
<tr>
<td>METEOR</td>
<td>.96</td>
<td>.88</td>
</tr>
<tr>
<td>TER</td>
<td>.83</td>
<td>.85</td>
</tr>
<tr>
<td>WER</td>
<td>.67</td>
<td>.83</td>
</tr>
<tr>
<td>TERRORCAT</td>
<td>.96</td>
<td>.95</td>
</tr>
<tr>
<td>DEPREF-ALIGN</td>
<td>.97</td>
<td>–</td>
</tr>
</tbody>
</table>

(System level, WMT 2013)
Correlations of metrics with human ranking

<table>
<thead>
<tr>
<th>Metric</th>
<th>de-en</th>
<th>en-de</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>.23</td>
<td>.18</td>
</tr>
<tr>
<td>METEOR</td>
<td>.26</td>
<td>.24</td>
</tr>
<tr>
<td>TERRORCAT</td>
<td>.25</td>
<td>.21</td>
</tr>
<tr>
<td>DEPREF-MAXCAT</td>
<td>.26</td>
<td>–</td>
</tr>
</tbody>
</table>

(Segment level, WMT 2013)
Automatic Metrics: Conclusions

- Automatic metrics essential tool for system development
- Not fully suited to rank systems of different types
- Reasonable results on system level evaluation, but not on sentence level
- Evaluation metrics still open challenge
For standard automatic metrics, a reference translation is needed.

In a translation scenario, we do not have reference translations.

It is very useful for a translator who is presented MT output to know:
- Is it good enough as it is
- Can it be easily edited
- Can it be edited with some effort
- Is it completely useless

This task is called quality estimation.
Quality Estimation – Details

- Automatic evaluation without a reference
- Typically modelled as a machine learning task
- Using features such as:
  - How long is the sentence?
  - What is the length difference between the source and target?
  - How common are the words and n-grams in the source sentence?
  - How ambiguous are the words in the source sentence?
  - How many punctuation marks are there in the sentence?
- Train on judgments of fluency/adequacy, post-editing effort, or post-editing time
Hypothesis Testing

- Situation
  - system A has score $x$ on a test set
  - system B has score $y$ on the same test set
  - $x > y$
- Is system A really better than system B?
- In other words:
  Is the difference in score statistically significant?
Core Concepts

- Null hypothesis
  - assumption that there is no real difference

- P-Levels
  - related to probability that there is a true difference
  - p-level $p < 0.01 = \text{more than } 99\% \text{ chance that difference is real}$
  - typically used: p-level 0.05 or 0.01

- Confidence Intervals
  - given that the measured score is $x$
  - what is the true score (on an infinite size test set)?
  - interval $[x - d, x + d]$ contains true score with, e.g., 95% probability
Pairwise Comparison

- Typically, we want to know if one system is better than another
  - Is system A better than system B?
  - Is change to my system an improvement?

- Example
  - Given a test set of 100 sentences
  - System A better on 60 sentences
  - System B better on 40 sentences

- Is system A really better?
Sign Test

Using binomial distribution

- system A better with probability $p_A$
- system B better with probability $p_B$ ($= 1 - p_A$)
- probability of system A better on $k$ sentences out of a sample of $n$ sentences

$$\binom{n}{k} p_A^k p_B^{n-k} = \frac{n!}{k!(n-k)!} p_A^k p_B^{n-k}$$

Null hypothesis: $p_A = p_B = 0.5$

$$\binom{n}{k} p^k (1 - p)^{n-k} = \binom{n}{k} 0.5^n = \frac{n!}{k!(n-k)!} 0.5^n$$
**Examples**

Given \( n \) sentences

system has to be better in at least \( k \) sentences

to achieve statistical significance at specified p-level

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p \leq 0.01 )</th>
<th>( p \leq 0.05 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>( k = 10 ) ( \frac{k}{n} = 1.00 )</td>
<td>( k \geq 9 ) ( \frac{k}{n} \geq 0.90 )</td>
</tr>
<tr>
<td>20</td>
<td>( k \geq 17 ) ( \frac{k}{n} \geq 0.85 )</td>
<td>( k \geq 15 ) ( \frac{k}{n} \geq 0.75 )</td>
</tr>
<tr>
<td>50</td>
<td>( k \geq 35 ) ( \frac{k}{n} \geq 0.70 )</td>
<td>( k \geq 33 ) ( \frac{k}{n} \geq 0.66 )</td>
</tr>
<tr>
<td>100</td>
<td>( k \geq 64 ) ( \frac{k}{n} \geq 0.64 )</td>
<td>( k \geq 61 ) ( \frac{k}{n} \geq 0.61 )</td>
</tr>
</tbody>
</table>
Data-driven Significance Testing

- Described methods require score at sentence level
- But: common metrics such as BLEU are computed for whole corpus
- Data-driven methods are typically used
- Bootstrap resampling
  - Sample sentences from the test set, with replacement
- Approximate randomization
  - Scramble sentences between the two systems that you compare
Summary

- MT evaluation is hard
- Human evaluation is expensive
- Automatic evaluation is cheap, but not always fair
- What is typically used in MT research:
  - Bleu!
  - Maybe another/several other metrics (typically Meteor, TER)
  - Maybe some human judgments
    - Ranking of systems
    - Targeted analysis of specific phenomenon
- → Be careful when you argue about MT quality!
Outlook

■ Next week: MT in practice
  ■ Guest lecture, Convertus (Commercial MT solutions in Uppsala)
  ■ Lab 1: Evaluation (Written lab report)
    **REMEMBER:** sign up for lab pairs!

■ Coming weeks:
  ■ Introduction to SMT
  ■ Lab 2: Word-based models
    ■ 1st part: oral examination in class
      be present and active for the full session
      (write report if absent/passive)
    ■ 2nd part: written lab report