Dr. Mariana Neves
(adapted from the original slides of Prof. Philipp Koehn)

December 5th, 2016
How good is a given machine translation system?

Hard problem, since many different translations are 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 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)
Evaluation metrics

- subjective judgments by human evaluators
- automatic evaluation metrics
  - how much post-editing effort?
  - does information come across?
Manual evaluation

- Bilingual evaluators are preferable, but not always available
- Alternatively, monolingual evaluators, if a reference translation is available
- Evaluation can be on sentence level, but the context should be checked as well.
Manual evaluation

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

Metrics

**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 English? 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>flawless English</td>
</tr>
<tr>
<td>4</td>
<td>good English</td>
</tr>
<tr>
<td>3</td>
<td>non-native English</td>
</tr>
<tr>
<td>2</td>
<td>disfluent English</td>
</tr>
<tr>
<td>1</td>
<td>incomprehensible</td>
</tr>
</tbody>
</table>

- **Adequacy**
  - 5: all meaning
  - 4: most meaning
  - 3: much meaning
  - 2: little meaning
  - 1: none

- **Fluency**
  - 5: flawless English
  - 4: good English
  - 3: non-native English
  - 2: disfluent English
  - 1: incomprehensible
**Judge Sentence**

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

**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>
<tr>
<td>both countries are a necessary laboratory at internal functioning of the eu .</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>the two countries are rather a laboratory necessary for the internal workings of the eu .</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>the two countries are rather a necessary laboratory internal workings of the eu .</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>

**Annotator:** Philipp Koehn  **Task:** WMT06 French-English

---

5= All Meaning  
4= Most Meaning  
3= Much Meaning  
2= Little Meaning  
1= None  
5= flawless English  
4= Good English  
3= Non-native English  
2= Disfluent English  
1= Incomprehensible
Evaluators Disagree

- Histogram of adequacy judgments by five 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
  (5-point scale → \( p(E) = \frac{1}{5} \))

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>
## Ranking Translations

- **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>
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 or to other domains
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
Israeli officials are responsible for airport security

**SYSTEM A:**

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

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

- **F-measure**
  
  \[
  \frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%
  \]
Israeli officials are responsible for airport security

<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
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>
BLEU (A Bilingual Evaluation Understudy)

- 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)

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

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

**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 (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>

Mariana Neves
Evaluation
December 5th, 2016
Multiple Reference Translations

- To account for variability, use multiple reference translations
  - n-grams may match in any of the references
  - the closest and smallest reference length is used
- Example

SYSTEM:

```
Israeli officials    responsibility of    airport   safety
```

REFERENCES:

```
Israel is in charge of the security at this airport
Israeli side was in charge of the security of this airport
```

The security work for this airport is the responsibility of the Israel government.
<table>
<thead>
<tr>
<th>Metric</th>
<th><strong>Single reference</strong></th>
<th></th>
<th><strong>Multiple reference</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>System A</td>
<td>System B</td>
<td>System A</td>
<td>System B</td>
</tr>
<tr>
<td>precision (1gram)</td>
<td>3/6</td>
<td>6/6</td>
<td>5/6</td>
<td>6/6</td>
</tr>
<tr>
<td>precision (2gram)</td>
<td>1/5</td>
<td>4/5</td>
<td>2/5</td>
<td>4/5</td>
</tr>
<tr>
<td>precision (3gram)</td>
<td>0/4</td>
<td>2/4</td>
<td>0/4</td>
<td>2/4</td>
</tr>
<tr>
<td>precision (4gram)</td>
<td>0/3</td>
<td>1/3</td>
<td>0/3</td>
<td>1/3</td>
</tr>
<tr>
<td>brevity penalty</td>
<td>6/7</td>
<td>6/7</td>
<td>6/7</td>
<td>6/7</td>
</tr>
<tr>
<td>BLEU-1</td>
<td>42%</td>
<td>86%</td>
<td>71%</td>
<td>86%</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>9%</td>
<td>69%</td>
<td>29%</td>
<td>69%</td>
</tr>
<tr>
<td>BLEU-3</td>
<td>0%</td>
<td>34%</td>
<td>0%</td>
<td>34%</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0%</td>
<td>11%</td>
<td>0%</td>
<td>11%</td>
</tr>
</tbody>
</table>
METEOR: Flexible Matching

- Partial credit for matching stems
  
  SYSTEM: Jim went home  
  REFERENCE: Joe goes home

- Partial credit for matching synonyms (e.g., Wordnet)
  
  SYSTEM: Jim walks home  
  REFERENCE: Joe goes home
Critique of Automatic Metrics

- They ignore relevance of words
  - names and core concepts more important than determiners and punctuation
  - negation cues (e.g., not) should not be missed
Critique of Automatic Metrics

- Operate on local level
  - do not consider overall grammar of the sentence
  - do not consider sentence meaning
  - it is supposed to favor phrase-based statistical systems
Critique of Automatic Metrics

- Scores are meaningless
  - very test-set specific, absolute value not informative
  - scores depend on many features:
    - number of reference translations
    - language pair
    - domain
    - tokenization scheme
Critique of Automatic Metrics

- Human translators score low on BLEU
  - possibly because of higher variability
  - different word choices
Automatic metrics are low cost, tunable, consistent

But are they correct?

→ Yes, if they correlate with human judgment
Correlation with Human Judgment (NIST’02, Arabic-English)

![Graph showing correlation between NIST Score (variant of BLEU) and Human Judgments]

- Adequacy:
- Fluency:

$R^2 = 90.2\%$
Pearson Correlation Coefficient

- Two variables: automatic score \( x \), human judgment \( y \)
- Multiple systems \((x_1, y_1), (x_2, y_2), \ldots\)
- Pearson's correlation coefficient \( r_{xy} \):
  \[
  r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \ s_x \ s_y}
  \]

- Note:
  \[
  \text{mean } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
  \]
  \[
  \text{variance } s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
  \]
### Pearson Correlation Coefficient

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>-0.29 to -0.10</td>
<td>0.10 to 0.20</td>
</tr>
<tr>
<td>medium</td>
<td>-0.49 to -0.30</td>
<td>0.30 to 0.49</td>
</tr>
<tr>
<td>high</td>
<td>-1.00 to -0.50</td>
<td>0.50 to 1.00</td>
</tr>
</tbody>
</table>
Evidence of Shortcomings of Automatic Metrics (NIST’05, Arabic-English)

Post-edited output vs. statistical systems

- Human Score
- Bleu Score
- Adequacy
- Correlation

Mariana Neves
Evaluation
December 5th, 2016
Evidence of Shortcomings of Automatic Metrics (NIST’05, Arabic-English)

Rule-based vs. statistical systems

- Human Score
- Bleu Score
- Adequacy
- Fluency

- Rule-based System (Systran)
- SMT System 1
- SMT System 2

Mariana Neves
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 correlation coefficient)
Automatic metrics are essential tools for system development

Not fully suited to rank systems of different types

Evaluation metrics still open challenge
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 \( p < 0.01 \) = more than 99\% 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 a infinite size test set)?
  - interval \([x - d, x + d]\) contains true score with, e.g., 95\% probability
Computing Confidence Intervals

- Example
  - 100 sentence translations evaluated
  - 30 found to be correct

True translation score?

(i.e. probability that any randomly chosen sentence is correctly translated)
true score lies in interval $[\bar{x} - d, \bar{x} + d]$ around sample score $\bar{x}$ with probability 0.95
Confidence Interval for Normal Distribution

- Compute mean $\bar{x}$ and variance $s^2$ from data of size $n$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- True mean $\mu$?
Student’s t-distribution

- Confidence interval $p(\mu \in [\bar{x} - d, \bar{x} + d]) \geq 0.95$ computed by

$$d = t \frac{s}{\sqrt{n}}$$

- Values for $t$ depend on test sample size and significance level:

<table>
<thead>
<tr>
<th>Significance Level</th>
<th>Test Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>99%</td>
<td>2.6259</td>
</tr>
<tr>
<td>95%</td>
<td>1.9849</td>
</tr>
<tr>
<td>90%</td>
<td>1.6602</td>
</tr>
</tbody>
</table>
Example

- **Given**
  - 100 sentence translations evaluated
  - 30 found to be correct

- **Sample statistics**
  - Sample mean $\bar{x} = \frac{30}{100} = 0.3$
  - Sample variance $s^2 = \frac{1}{99} (70 \times (0 - 0.3)^2 + 30 \times (1 - 0.3)^2) = 0.2121$

- Consulting table for $t$ with 95% significance $\rightarrow 1.9849$

- Computing interval $d = 1.9849 \times \frac{0.2121}{\sqrt{100}} = 0.042 \rightarrow [0.258; 0.342]$
Pairwise Comparison

- Typically, absolute score less interesting
- More important
  - 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

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p \leq 0.01$</th>
<th>$p \leq 0.05$</th>
<th>$p \leq 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>$k \geq 10$</td>
<td>$k \geq 9$</td>
<td>$k = 5$</td>
</tr>
<tr>
<td></td>
<td>$\frac{k}{n} \geq 1.00$</td>
<td>$\frac{k}{n} \geq 0.90$</td>
<td>$\frac{k}{n} \geq 1.00$</td>
</tr>
<tr>
<td>10</td>
<td>$k \geq 17$</td>
<td>$k \geq 15$</td>
<td>$k \geq 15$</td>
</tr>
<tr>
<td></td>
<td>$\frac{k}{n} \geq 0.85$</td>
<td>$\frac{k}{n} \geq 0.75$</td>
<td>$\frac{k}{n} \geq 0.75$</td>
</tr>
<tr>
<td>50</td>
<td>$k \geq 35$</td>
<td>$k \geq 33$</td>
<td>$k \geq 32$</td>
</tr>
<tr>
<td></td>
<td>$\frac{k}{n} \geq 0.70$</td>
<td>$\frac{k}{n} \geq 0.66$</td>
<td>$\frac{k}{n} \geq 0.64$</td>
</tr>
<tr>
<td>100</td>
<td>$k \geq 64$</td>
<td>$k \geq 61$</td>
<td>$k \geq 59$</td>
</tr>
<tr>
<td></td>
<td>$\frac{k}{n} \geq 0.64$</td>
<td>$\frac{k}{n} \geq 0.61$</td>
<td>$\frac{k}{n} \geq 0.59$</td>
</tr>
</tbody>
</table>

Given $n$ sentences, the system has to be better in at least $k$ sentences to achieve statistical significance at specified p-level.
Sampling

Described methods require score at sentence level
But: common metrics such as BLEU are computed for whole corpus

Sampling

1. test set of 2000 sentences, sampled from large collection
2. compute the BLEU score for this set
3. repeat step 1–2 for 1000 times
4. ignore 25 highest and 25 lowest obtained BLEU scores
→ 95% confidence interval
Bootstrap resampling: sample from the same 2000 sentences, with replacement.

For pairwise comparison, which check which system is better on each sample. If a system is better on at least 950 samples, it is better under a statistically significance of $p \leq 0.05$. 

Mariana Neves
Evaluation
December 5th, 2016 48 / 52
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

But: time consuming, depend on skills of translator and post-editor

Metrics inspired by this task
- TER (translation edit rate): based on number of editing steps (Levenshtein operations: insertion, deletion, substitution) plus movement
- HTER (human translation edit rate): manually construct reference translation for output, apply TER (very time consuming, used in DARPA GALE program 2005-2011)
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

- Evaluation is based on the ratio of correctly answered questions and the time spent
Suggested reading

- Statistical Machine Translation, Philipp Koehn (chapter 8).