MT Evaluation: Human Measures and Assessment Methods

11-731:
Machine Translation
Alon Lavie
February 12, 2013
Need for MT Evaluation

• MT Evaluation is important:
  – MT systems are becoming wide-spread, embedded in more complex systems
    • How well do they work in practice?
    • Are they reliable enough?
  – MT is a technology still in research stages
    • How can we tell if we are making progress?
    • Metrics that can drive experimental development

• MT Evaluation is difficult:
  – Language Variability: there is no single correct translation
  – Human evaluation is subjective
  – How good is “good enough”? Depends on target application
  – Is system A better than system B? Depends on specific criteria...

• MT Evaluation is a research topic in itself! How do we assess whether an evaluation method is good?
Dimensions of MT Evaluation

- Human evaluation vs. automatic metrics
- Quality assessment at sentence (segment) level vs. system-level vs. task-based evaluation
- “Black-box” vs. “Glass-box” evaluation
- Evaluation for external validation vs. contrastive comparison of different MT systems vs. target function for automatic MT system tuning
Human Evaluation of MT Output

Why perform human evaluation?

- **Automatic MT metrics are not sufficient:**
  - What does a BLEU score of 30.0 or 50.0 mean?
  - Existing automatic metrics are rather crude and at times biased
  - Automatic metrics usually don’t provide sufficient insight for error analysis
  - Different types of errors have different implications depending on the underlying task in which MT is used

- **Need for reliable human measures in order to develop and assess automatic metrics for MT evaluation**
Human Evaluation: Main Challenges

• Time and Cost
• Reliability and Consistency: difficulty in obtaining high-levels of intra and inter-coder agreement
  – **Intra-coder Agreement**: consistency of same human judge
  – **Inter-coder Agreement**: judgment agreement across multiple judges of quality
• Measuring Reliability and Consistency
• Developing meaningful metrics based on collected human judgments
  – Example: if collecting binary judgments for sentences, how do these map into global scores?
Main Types of Human Assessments

- Adequacy and Fluency scores
- Human ranking of translations at the sentence-level
- Post-editing Measures:
  - Post-editor editing time/effort measures
  - HTER: Human Translation Edit Rate
- Human Post-Editing measures: can humans edit the MT output into a correct translation?
- Task-based evaluations: was the performance of the MT system sufficient to perform a particular task?
Adequacy and Fluency

- **Adequacy**: is the meaning translated correctly?
  - By comparing MT translation to a reference translation (or to the source)?
- **Fluency**: is the output grammatical and fluent?
  - By comparing MT translation to a reference translation, to the source, or in isolation?
- **Scales**: [1-5], [1-10], [1-7]
- **Initiated during DARPA MT evaluations during mid-1990s**
- **Most commonly used until recently**
- **Main Issues**: definitions of scales, agreement, normalization across judges
Human Ranking of MT Output

• Method: compare two or more translations of the same sentence and rank them in quality
  – More intuitive, less need to define exact criteria
  – Can be problematic: comparing bad long translations is very confusing and unreliable

• Main Issues:
  – Binary rankings or multiple translations?
  – Agreement levels
  – How to use ranking scores to assess systems?
Human Assessment in WMT-2012

- WMT-2012: Shared task on developing MT systems between several European languages (to English and from English)
- Also included tracks on automated MT metric evaluation and quality estimation
- Official Metric: Human Rankings
- Detailed evaluation and analysis of results
- 2-day Workshop at NAACL-2012, including detailed analysis paper by organizers
Human Rankings at WMT-2012

- **Instructions:** Rank translations from Best to Worst relative to the other choices (ties are allowed)
- Annotators were shown at most five translations at a time.
- For all language pairs there were more than 5 system submissions. No attempt to get a complete ordering over all the systems at once.
- Relied on random selection and a reasonably large sample size to make the comparisons fair.
- **Metric to compare MT systems:** Individual systems are ranked based on the fraction of comparison instances for which they were judged to be better than any other system.
Assessing MT Systems

• Human Rankings were used to assess:
  – Which systems produced the best translation quality for each language pair?
  – Which of the systems that used only the provided training materials ("constrained") produced the best translation quality?
<table>
<thead>
<tr>
<th>System</th>
<th>C7</th>
<th>&gt;-others</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONLINE-B</td>
<td>Y</td>
<td>0.65</td>
</tr>
<tr>
<td>UEDIN</td>
<td>Y</td>
<td>0.60</td>
</tr>
<tr>
<td>CU-SOAR</td>
<td>Y</td>
<td>0.53</td>
</tr>
<tr>
<td>ONLINE-A</td>
<td>N</td>
<td>0.53</td>
</tr>
<tr>
<td>UK</td>
<td>Y</td>
<td>0.37</td>
</tr>
<tr>
<td>JHU</td>
<td>Y</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Spanish-English

<table>
<thead>
<tr>
<th>System</th>
<th>C7</th>
<th>&gt;-others</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONLINE-B</td>
<td>Y</td>
<td>0.62</td>
</tr>
<tr>
<td>UEDIN</td>
<td>Y</td>
<td>0.61</td>
</tr>
<tr>
<td>GOO</td>
<td>Y</td>
<td>0.60</td>
</tr>
<tr>
<td>UEDIN</td>
<td>Y</td>
<td>0.57</td>
</tr>
<tr>
<td>UPC</td>
<td>Y</td>
<td>0.52</td>
</tr>
<tr>
<td>CITH-UTM</td>
<td>Y</td>
<td>0.52</td>
</tr>
<tr>
<td>JHU</td>
<td>Y</td>
<td>0.48</td>
</tr>
<tr>
<td>KASMT-3</td>
<td>N</td>
<td>0.66</td>
</tr>
<tr>
<td>KASMT-1</td>
<td>N</td>
<td>0.62</td>
</tr>
<tr>
<td>ONLINE-C</td>
<td>N</td>
<td>0.42</td>
</tr>
<tr>
<td>UK</td>
<td>Y</td>
<td>0.19</td>
</tr>
</tbody>
</table>

French-English

<table>
<thead>
<tr>
<th>System</th>
<th>C7</th>
<th>&gt;-others</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONLINE-A</td>
<td>N</td>
<td>0.59</td>
</tr>
<tr>
<td>CMU</td>
<td>Y</td>
<td>0.57</td>
</tr>
<tr>
<td>ONLINE-B</td>
<td>N</td>
<td>0.57</td>
</tr>
<tr>
<td>UEDIN</td>
<td>Y</td>
<td>0.55</td>
</tr>
<tr>
<td>L1LM</td>
<td>Y</td>
<td>0.52</td>
</tr>
<tr>
<td>RNTT</td>
<td>Y</td>
<td>0.52</td>
</tr>
<tr>
<td>KASMT-3</td>
<td>N</td>
<td>0.46</td>
</tr>
<tr>
<td>KASMT-1</td>
<td>N</td>
<td>0.33</td>
</tr>
<tr>
<td>JHU</td>
<td>Y</td>
<td>0.44</td>
</tr>
<tr>
<td>ONLINE-C</td>
<td>N</td>
<td>0.32</td>
</tr>
</tbody>
</table>

English-French

<table>
<thead>
<tr>
<th>System</th>
<th>C7</th>
<th>&gt;-others</th>
</tr>
</thead>
<tbody>
<tr>
<td>UEDIN</td>
<td>Y</td>
<td>0.66</td>
</tr>
<tr>
<td>KASMT-3</td>
<td>N</td>
<td>0.62</td>
</tr>
<tr>
<td>RNTT</td>
<td>Y</td>
<td>0.59</td>
</tr>
<tr>
<td>L1LM</td>
<td>Y</td>
<td>0.52</td>
</tr>
<tr>
<td>UEDIN</td>
<td>Y</td>
<td>0.53</td>
</tr>
<tr>
<td>KASMT-3</td>
<td>N</td>
<td>0.51</td>
</tr>
<tr>
<td>KASMT-1</td>
<td>N</td>
<td>0.48</td>
</tr>
<tr>
<td>JHU</td>
<td>Y</td>
<td>0.44</td>
</tr>
<tr>
<td>ONLINE-C</td>
<td>N</td>
<td>0.39</td>
</tr>
<tr>
<td>ITIS-LATI</td>
<td>N</td>
<td>0.36</td>
</tr>
</tbody>
</table>

C7 indicates whether system is constrained (unhighlighted rows): trained only using supplied training data, standard monolingual linguistic tools, and, optionally, LDC’s English Gigaword.

* indicates a win: no other system is statistically significantly better at p-level < 0.10 in pairwise comparison.

indicates a constrained win: no other constrained system is statistically better.

Table 4: Official results for the WMT12 translation task. Systems are ordered by their >-others score, reflecting how often their translations won in pairwise comparisons. For detailed head-to-head comparisons, see Appendix A.
### French-English
1,437–1,701 comparisons/system

<table>
<thead>
<tr>
<th>System</th>
<th>C?</th>
<th>&gt;others</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIMSI ⋆⋆</td>
<td>Y</td>
<td>0.63</td>
</tr>
<tr>
<td>KIT ⋆⋆</td>
<td>Y</td>
<td>0.61</td>
</tr>
<tr>
<td>ONLINE-A ⋆</td>
<td>N</td>
<td>0.59</td>
</tr>
<tr>
<td>CMU ⋆⋆</td>
<td>Y</td>
<td>0.57</td>
</tr>
<tr>
<td>ONLINE-B ⋆</td>
<td>N</td>
<td>0.57</td>
</tr>
<tr>
<td>UEDIN</td>
<td>Y</td>
<td>0.55</td>
</tr>
<tr>
<td>LIUM</td>
<td>Y</td>
<td>0.52</td>
</tr>
<tr>
<td>RWTH</td>
<td>Y</td>
<td>0.52</td>
</tr>
<tr>
<td>RBMT-1</td>
<td>N</td>
<td>0.46</td>
</tr>
<tr>
<td>RBMT-3</td>
<td>N</td>
<td>0.46</td>
</tr>
<tr>
<td>UK</td>
<td>Y</td>
<td>0.44</td>
</tr>
<tr>
<td>SFU</td>
<td>Y</td>
<td>0.44</td>
</tr>
<tr>
<td>RBMT-4</td>
<td>N</td>
<td>0.43</td>
</tr>
<tr>
<td>JHU</td>
<td>Y</td>
<td>0.41</td>
</tr>
<tr>
<td>ONLINE-C</td>
<td>N</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Methods for Overall Ranking

- Different possible ways to calculate overall system rankings based on the collected segment-level ranking judgments
- WMT-2012 surveys six different possible methods and compares five of them on the data collected for English-German MT systems
- Different methods generate mostly but not fully similar results
- Statistical significance can be established based on the variance within the collected data, using bootstrap sampling
## Methods for Overall Ranking

<table>
<thead>
<tr>
<th>Bojar</th>
<th>Lopez</th>
<th>Most Probable</th>
<th>MC Playoffs</th>
<th>Expected Wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.641: ONLINE-B</td>
<td>RBMT-4</td>
<td>RBMT-4</td>
<td>6.16: ONLINE-B</td>
</tr>
<tr>
<td>2</td>
<td>0.627: RBMT-3</td>
<td>ONLINE-B</td>
<td>ONLINE-B</td>
<td>6.39: RBMT-3</td>
</tr>
<tr>
<td>3</td>
<td>0.577: RBMT-4</td>
<td>RBMT-3</td>
<td>RBMT-3</td>
<td>6.98: RBMT-4</td>
</tr>
<tr>
<td>4</td>
<td>0.557: RBMT-1</td>
<td>RBMT-1</td>
<td>RBMT-1</td>
<td>7.32: RBMT-1</td>
</tr>
<tr>
<td>5</td>
<td>0.547: LIMSI</td>
<td>ONLINE-A</td>
<td>ONLINE-A</td>
<td>7.46: LIMSI</td>
</tr>
<tr>
<td>6</td>
<td>0.537: ONLINE-A</td>
<td>UEDIN-WILLIAMS</td>
<td>LIMSI</td>
<td>7.57: ONLINE-A</td>
</tr>
<tr>
<td>7</td>
<td>0.509: UEDIN-WILLIAMS</td>
<td>LIMSI</td>
<td>UEDIN-WILLIAMS</td>
<td>7.87: UEDIN-WILLIAMS</td>
</tr>
<tr>
<td>8</td>
<td>0.503: KIT</td>
<td>KIT</td>
<td>KIT</td>
<td>7.98: KIT</td>
</tr>
<tr>
<td>9</td>
<td>0.476: DFKI-HUNSICKER</td>
<td>DFKI-HUNSICKER</td>
<td>DFKI-HUNSICKER</td>
<td>8.32: DFKI-HUNSICKER</td>
</tr>
<tr>
<td>10</td>
<td>0.475: UEDIN</td>
<td>ONLINE-C</td>
<td>ONLINE-C</td>
<td>8.38: DFKI-HUNSICKER</td>
</tr>
<tr>
<td>11</td>
<td>0.470: RWTH</td>
<td>UEDIN</td>
<td>UEDIN</td>
<td>8.41: ONLINE-C</td>
</tr>
<tr>
<td>12</td>
<td>0.470: ONLINE-C</td>
<td>UK</td>
<td>UK</td>
<td>8.44: RWTH</td>
</tr>
<tr>
<td>13</td>
<td>0.448: UK</td>
<td>RWTH</td>
<td>RWTH</td>
<td>8.72: UK</td>
</tr>
<tr>
<td>14</td>
<td>0.435: JIU</td>
<td>JIU</td>
<td>JIU</td>
<td>8.87: JIU</td>
</tr>
<tr>
<td>15</td>
<td>0.249: DFKI-BERLIN</td>
<td>DFKI-BERLIN</td>
<td>DFKI-BERLIN</td>
<td>11.15: DFKI-BERLIN</td>
</tr>
</tbody>
</table>

Table 5: Overall ranking with different methods (English–German)
Human Post-Editing

- A natural task-based evaluation measure for utility of MT output
  - Human translator(s) edit the output of the MT system into a correct translation
  - Measure the amount of “effort” involved
- Practical: increasing number of commercial translation agencies are actually doing MTPE
- Challenges:
  - How do you measure post-editing “effort”?  
  - Large variations across translators – training is important
  - Bilingual translators are costly – can monolingual target-language speakers do this reliably?
**TER**

- Translation Edit (Error) Rate (Snover et. al. 2006)
- Main Ideas:
  - Edit-based measure, similar in concept to Levenshtein distance: counts the number of word **insertions, deletions and substitutions** required to transform the MT output to the reference translation
  - Adds the notion of “block movements” as a single edit operation
  - Only **exact word matches** count, but latest version (TERp) incorporates synonymy and paraphrase matching and tunable parameters
  - Can be used as a rough post-editing measure, but is not a true measure of post-editing effort
**HTER**

- **Human Translation Edit Rate**
- Developed as the official evaluation measure of the DARPA GALE program and continues to be used in BOLT

**Evaluation Process:**
- Team of translators post-edits the MT segment
- TER is used to find the minimum-distance post-edited human reference
- Aggregate system-level HTER scores are calculated at the document-level
- Ranked document lists are generated for each system
- Systems are scored based on fraction of documents that pass threshold levels of TER performance
Human Editing at WMT-2009

- Two Stages:
  - Humans edit the MT output to make it as fluent as possible
  - Judges evaluate the **edited output** for adequacy (meaning) with a **binary** Y/N judgment

- Instructions:
  - **Step-1:** Correct the translation displayed, making it as fluent as possible. If no corrections are needed, select “No corrections needed.” If you cannot understand the sentence well enough to correct it, select “Unable to correct.”
  - **Step-2:** Indicate whether the edited translations represent fully fluent and meaning equivalent alternatives to the reference sentence. The reference is shown with context, the actual sentence is bold.
Editing Interface

**WMT09 Manual Evaluation**

**Edit MT Output**

You have judged 19 sentences for WMT09 Multisource-English News Editing, 468 sentences total taking 74.4 seconds per sentence.

Original: They are often linked to other alterations, such as nightmares, night terrors, the nocturnal enuresis (bedwetting), or the sleepwalking, but it is not always the case.

**Edit:**
They are often linked to other sleep disorders, such as nightmares, night terrors, the nocturnal enuresis (bedwetting) or sleepwalking, but this is not always the case.

- Reset Edit
- @ Edited
- ◼️ No corrections needed
- ◼️ Unable to correct

**Annotator:** ccb Task: WMT09 Multisource-English News Editing

**Instructions:**
Correct the translation displayed, making it as fluent as possible. If no corrections are needed, select "No corrections needed." If you cannot understand the sentence well enough to correct it, select "Unable to correct."

Figure 2: This screenshot shows an annotator editing the output of a machine translation system.
Evaluating Edited Output

Figure 3: This screenshot shows an annotator judging the acceptability of edited translations.
Human Editing Results

- **Goal**: to assess how often a systems MT output is “fixable” by a human post-editor
- **Measure used**: fraction of time that humans assessed that the edited output had the same meaning as the reference
Assessing Coding Agreement

• **Intra**-annotator Agreement:
  – 10% of the items were repeated and evaluated twice by each judge.

• **Inter**-annotator Agreement:
  – 40% of the items were randomly drawn from a common pool that was shared across all annotators creating a set of items that were judged by multiple annotators.

• Agreement Measure: Kappa Coefficient

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

- $P(A)$ is the proportion of times that the annotators agree
- $P(E)$ is the proportion of time that they would agree by chance.
Assessing Coding Agreement

Common Interpretation of Kappa Values:
0.0-0.2: slight agreement
0.2-0.4: fair agreement
0.4-0.6: moderate agreement
0.6-0.8: substantial agreement
0.8-1.0: near perfect agreement

Table 3: Inter- and intra-annotator agreement rates for the WMT12 manual evaluation. For comparison, the WMT11 rows contain the results from the European languages individual systems task (Callison-Burch et al. (2011), Table 7).
Figure 4: The effect of discarding every annotators’ initial judgments, up to the first 50 items.

Figure 5: The effect of removing annotators with the lowest agreement, disregarding up to 40 annotators.
Normalizing Human Bias

• Human judgments using absolute scales (Likert Scores) typically exhibit subjective biases among judges
• Normalizing scores across judges can significantly improve inter-coder agreement
• Several normalization methods have been proposed in recent years
• One example: (Blatz et al. 2003)
  - Normalize the scores into a continuous space [0-1] by mapping each discrete score $s$ to the fraction of judgments of score $\leq s$
Cost and Quality Issues

- **High cost** and controlling for agreement quality are the most challenging issues in conducting human evaluations of MT output.
- Critical decisions:
  - Your human judges: professional translators? Non-expert bilingual speakers? Target-language only speakers?
  - Where do you recruit them? How do you train them?
  - How many different judgments per segment to collect?
  - Easy to overlook issues (i.e. the user interface) can have significant impact on quality and agreement.
- Measure intra- and inter-coder agreement as an integral part of your evaluation!
Human Evaluations Using Crowd-Sourcing

- Recent popularity of crowd-sourcing has introduced some exciting new ideas for human assessment of MT output
  - Using the “crowd” to provide human judgments of MT quality, either directly or indirectly
  - Amazon’s Mechanical Turk as a labor source for human evaluation of MT output
Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

56,611 HITs available. View them now.

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Get started.

As a Mechanical Turk Requester you:
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results
**Mechanical Turk**

![Mechanical Turk](image)

**February 12, 2013**

11-731: Machine Translation
<table>
<thead>
<tr>
<th>Rate this translation</th>
<th>Human Translation</th>
<th>Automatic Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Arabic)</td>
<td>(Arabic)</td>
<td>(Arabic)</td>
</tr>
<tr>
<td>4 - Excellent</td>
<td>4 - Excellent</td>
<td>4 - Excellent</td>
</tr>
<tr>
<td>3 - Good</td>
<td>3 - Good</td>
<td>3 - Good</td>
</tr>
<tr>
<td>2 - Bad</td>
<td>2 - Bad</td>
<td>2 - Bad</td>
</tr>
<tr>
<td>1 - Very bad</td>
<td>1 - Very bad</td>
<td>1 - Very bad</td>
</tr>
</tbody>
</table>

**Instructions (English):** Below are two translations of the same English sentence into Arabic. The first was written by a human translator and the second was translated automatically by a computer. Please rate the extent to which the automatic translation has the same meaning as the human translation.

**Rate:****  **

<table>
<thead>
<tr>
<th>Human translation</th>
<th>4 - Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 - Good</td>
<td>2 - Bad</td>
</tr>
<tr>
<td>1 - Very bad</td>
<td>1 - Very bad</td>
</tr>
<tr>
<td>Automatic translation</td>
<td>4 - Excellent</td>
</tr>
<tr>
<td>3 - Good</td>
<td>2 - Bad</td>
</tr>
<tr>
<td>1 - Very bad</td>
<td>1 - Very bad</td>
</tr>
</tbody>
</table>

**Scale and Examples:**

1. **Good ( translating into Arabic)**
   - في هذه الصناعة ذات القوة العالمية سككون 328 مليار دولار، بالإضافة إلى إيرادات أخرى تصل إلى 100 مليار دولار، والفرصة الربحية.
   - معكوساً، النجاح العربي في مجال تدريس اللغة العربية، هما

---

February 12, 2013 11-731: Machine Translation
Summary

• Human assessment of MT output is still extremely important... even though it is difficult to do reliably, and there is no clear consensus on best practice methods

• Human and automatic metrics are both essential in modern MT development and serve different purposes

• Good human metrics greatly help in developing good automatic metrics
Questions?