(Meta-) Evaluation of Machine Translation

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, Josh Schroeder

June 23, 2007
Shared Task Evaluation

- Shared translation task:
  - English to/from French, German, Spanish, and Czech
  - Test sets drawn from Europarl and news commentary
- 88 ‘primary’ system submissions were manually evaluated
- Recruited 100+ judges, who contributed 330 hours for 81,000+ judgments
- Stated Goal: Determine system performance, establish reference for future research
- Ulterior Motive
The “Meta” Part

• We wanted to analyze evaluation measures and establish “best practices”

• Questions to investigate:
  – Which automatic evaluation metrics correlate most strongly with human judgments of translation quality?
  – How consistent are people when they judge translation quality?
  – To what extent do they agree with other annotators?
  – Can we improve human evaluation?

• Structured evaluation so that we could answer these
Manual Evaluation

- Three different types of evaluation:
  - Score each translation along *fluency* and *adequacy* scales
  - Rank translations of sentences relative to each other
  - Rank translations of sub-sentential units

- Meta-evaluation measured:
  - Inter-annotator agreement (agreement with others)
  - Intra-annotator agreement (self consistency)
  - Average time to make one judgement
Ranking Translations of Constituents

- Intuition: Ranking translations of long sentences is difficult, because systems produces errors in different parts of them
- Goal: focus attention on particular parts of the translation to make the task easier
- Method:
  1. Automatically align source with reference and system translations
  2. Parse source sentence
  3. Select constituents to be judged
  4. Highlight source phrase and corresponding target phrases
  5. Rank those
<table>
<thead>
<tr>
<th>people</th>
<th>Iraq</th>
<th>services</th>
<th>provide</th>
<th>basic</th>
<th>cannot</th>
<th>sustain</th>
<th>occupation</th>
<th>its</th>
<th>the</th>
<th>Can</th>
</tr>
</thead>
<tbody>
<tr>
<td>König</td>
<td>die</td>
<td>USA</td>
<td>ihre</td>
<td>Besetzung</td>
<td>aufrechterhalten</td>
<td>wenn</td>
<td>sie</td>
<td>dem</td>
<td>irakischen</td>
<td>Volk</td>
</tr>
</tbody>
</table>

**Note:** The text appears to be a table with German words, possibly indicating a comparison or classification of some sort. The table lacks context or a clear title, making it difficult to interpret without additional information.
People's basic services cannot be provided to Iraq's people if its occupation by the USA is sustained.
Can the US sustain its occupation if it cannot provide basic services to Iraq's people?
The US can maintain its occupation of Iraq to provide basic services to people's sustain its occupation cannot sustain it. It is when the Besetzung aufrechterhalten ihre Dienstleistungen anbieten können.
The US can maintain its occupation if they cannot offer the Iraqi people basic services.
können Dienstleistungen grundlegende Volk irakischen Besetzung aufrechterhalten können

The United States can maintain its employment when it the Iraqi people not basic services on offer

Can the US sustain its occupation if it cannot provide basic services to Iraq 's people
The United States can maintain its occupation when it cannot sustain its basic services to the Iraqi people. The occupation continues even though the US cannot provide basic services to the Iraqi people.
Results of the Meta-Evaluation

- We measured agreement among annotators using the kappa coefficient:

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

where
- \( P(A) \) is the proportion of times that the annotators agree
- \( P(E) \) is the proportion of time that they would agree by chance.

- Interpretation of \( K \) scores varies, but:
  - \(.6 \pm .8\) is good agreement
  - \(.4 \pm .6\) is moderate agreement
  - \(< .4\) and we should start to worry
## Inter-Annnotator Agreement

<table>
<thead>
<tr>
<th>Evaluation type</th>
<th>$P(A)$</th>
<th>$P(E)$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency (absolute)</td>
<td>.400</td>
<td>.2</td>
<td>.250</td>
</tr>
<tr>
<td>Adequacy (absolute)</td>
<td>.380</td>
<td>.2</td>
<td>.226</td>
</tr>
<tr>
<td>Fluency (relative)</td>
<td>.520</td>
<td>.333</td>
<td>.281</td>
</tr>
<tr>
<td>Adequacy (relative)</td>
<td>.538</td>
<td>.333</td>
<td>.307</td>
</tr>
<tr>
<td>Sentence ranking</td>
<td>.582</td>
<td>.333</td>
<td>.373</td>
</tr>
<tr>
<td>Constituent ranking</td>
<td>.712</td>
<td>.333</td>
<td>.566</td>
</tr>
</tbody>
</table>
## Intra-Annotator Agreement

<table>
<thead>
<tr>
<th>Evaluation type</th>
<th>$P(A)$</th>
<th>$P(E)$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency (absolute)</td>
<td>.630</td>
<td>.2</td>
<td>.537</td>
</tr>
<tr>
<td>Adequacy (absolute)</td>
<td>.574</td>
<td>.2</td>
<td>.468</td>
</tr>
<tr>
<td>Fluency (relative)</td>
<td>.690</td>
<td>.333</td>
<td>.535</td>
</tr>
<tr>
<td>Adequacy (relative)</td>
<td>.696</td>
<td>.333</td>
<td>.544</td>
</tr>
<tr>
<td>Sentence ranking</td>
<td>.749</td>
<td>.333</td>
<td>.623</td>
</tr>
<tr>
<td>Constituent ranking</td>
<td>.842</td>
<td>.333</td>
<td>.762</td>
</tr>
</tbody>
</table>
Time to judge one item

- Constituent rank
- Sentence rank
- Fluency+adequacy scoring

Number of items taking this long (%) vs. time to judge one item (seconds)
Automatic evaluation metrics

- Ranked system outputs using 11 different automatic metrics
  N-gram matching:
    *Bleu, GTM, Translation Error Rate*
  Flexible matching:
    *Meteor, ParaEval precision, ParaEval recall*
  Linguistic info:
    *Dependency overlap, Semantic role overlap, WER over verbs*
  Correlation-centric:
    *Maximum correlation training on adequacy, and on fluency*
- Meta-evaluation: Spearman’s rank correlation with human judgments
Proportion of time entries were top-ranked in manual evaluation

<table>
<thead>
<tr>
<th>Institution</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYSTRAN</td>
<td>32%</td>
</tr>
<tr>
<td>University of Edinburgh</td>
<td>20%</td>
</tr>
<tr>
<td>University of Catalonia</td>
<td>15%</td>
</tr>
<tr>
<td>LIMSI-CNRS</td>
<td>13%</td>
</tr>
<tr>
<td>University of Maryland</td>
<td>5%</td>
</tr>
<tr>
<td>National Research Council + SYSTRAN</td>
<td>5%</td>
</tr>
<tr>
<td>Commercial Czech-English system</td>
<td>5%</td>
</tr>
<tr>
<td>University of Valencia</td>
<td>2%</td>
</tr>
<tr>
<td>Charles University</td>
<td>2%</td>
</tr>
</tbody>
</table>
Proportion of time entries were top-ranked by automatic metrics

University of Edinburgh  41%
University of Catalonia  12%
LIMSI-CNRS  12%
University of Maryland  9%
Charles University  4%
Carnegie Mellon University  4%
Carnegie Mellon University  4%
University of California at Berkeley  3%
National Research Council + SYSTRAN  2%
SYSTRAN  2%
Saarland University  0.8%
## Correlation

<table>
<thead>
<tr>
<th></th>
<th>Adequacy</th>
<th>Fluency</th>
<th>Rank</th>
<th>Constituent</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic role</td>
<td>.77</td>
<td>.84</td>
<td>.80</td>
<td>.74</td>
<td>.79</td>
</tr>
<tr>
<td>ParaEval-Recall</td>
<td>.71</td>
<td>.74</td>
<td>.77</td>
<td>.80</td>
<td>.76</td>
</tr>
<tr>
<td>Meteor</td>
<td>.71</td>
<td>.72</td>
<td>.75</td>
<td>.67</td>
<td>.71</td>
</tr>
<tr>
<td><em>Bleu</em></td>
<td>.69</td>
<td>.72</td>
<td>.67</td>
<td>.60</td>
<td>.67</td>
</tr>
<tr>
<td>Max adeq corr</td>
<td>.65</td>
<td>.66</td>
<td>.66</td>
<td>.53</td>
<td>.63</td>
</tr>
<tr>
<td>Max flu corr</td>
<td>.64</td>
<td>.65</td>
<td>.66</td>
<td>.51</td>
<td>.61</td>
</tr>
<tr>
<td>GTM</td>
<td>.66</td>
<td>.67</td>
<td>.62</td>
<td>.50</td>
<td>.61</td>
</tr>
<tr>
<td>Dependency overlap</td>
<td>.64</td>
<td>.64</td>
<td>.60</td>
<td>.51</td>
<td>.60</td>
</tr>
<tr>
<td>ParaEval-Precision</td>
<td>.64</td>
<td>.65</td>
<td>.61</td>
<td>.49</td>
<td>.60</td>
</tr>
<tr>
<td>1-TER</td>
<td>.61</td>
<td>.54</td>
<td>.52</td>
<td>.51</td>
<td>.54</td>
</tr>
<tr>
<td>1-WER of verbs</td>
<td>.38</td>
<td>.42</td>
<td>.43</td>
<td>.30</td>
<td>.38</td>
</tr>
</tbody>
</table>
Reflections

- Should be careful to evaluate rigorously when making scientific claims
- Seemed to be a consistent bias in automatic metrics
- Agreement was low for fluency and adequacy scores
- Suggests to me that
  1. Should avoid depending on automatic metrics in conference papers
  2. Should research ways of improving manual evaluation so that it is
     - more consistent
     - faster / cheaper
     - easier to perform
     - re-usable
Future Evaluations

• Euromatrix project will be hosting an ongoing online evaluation

• Goals:
  – Provide common test sets and training data,
  – Provide means for asynchronous evaluation
  – Collect translations, show off best of best

• Expanded in scope to translation between all 23 official European languages
  – That’s 242 language pairs, and 484 directions!
  – You could have the best Latvian-Maltese translation system in the world!

• Continue annual evaluation, which will focus on a subset of languages and do extensive manual evaluation
Thank You!

- Questions?
- Segue into “poster boaster” session
Best German-English Systems

- German → English Europarl:
  \[\text{SYSTRAN} > \text{liu} > \text{uedin} = \text{upc} > \text{cmu-uka} > \text{nrc} > \text{saar}\]

- German → English News Corpus:
  \[\text{SYSTRAN} > \text{uedin} > \text{upc} > \text{nrc} > \text{saar}\]

- English → German Europarl:
  \[\text{UEDIN} > \text{systran} = \text{upc} > \text{cmu-uka} > \text{nrc} > \text{saar}\]

- English → German News Corpus:
  \[\text{SYSTRAN} > \text{upc} > \text{uedin} > \text{nrc} > \text{ucb} > \text{saar}\]
Best Spanish-English Systems

- Spanish → English Europarl:
  \[ \text{UPC} = \text{UEDIN} > \text{upv} > \text{cmu-synta}x > \text{cmu-uka} = \text{systran} > \text{nc} > \text{saar} \]

- Spanish → English News Corpus:
  \[ \text{UPC} > \text{uedin} > \text{systran} > \text{cmu-uka} > \text{nc} > \text{upv} > \text{saar} \]

- English → Spanish Europarl:
  \[ \text{UEDIN} > \text{upc} = \text{upv} > \text{cmu-uka} > \text{nc} = \text{systran} \]

- English → Spanish News Corpus:
  \[ \text{SYSTRAN} > \text{upc} > \text{cmu-uka} > \text{ucb} > \text{uedin} > \text{nc} = \text{upv} \]
Best French-English Systems

- **French → English Europarl:**
  
  \[
  \text{LIMSI} = \text{UEDIN} > \text{systran-nrc} = \text{upc} > \text{nrc} > \text{systran} > \text{saar}
  \]

- **French → English News Corpus:**
  
  \[
  \text{LIMSI} > \text{upc} = \text{uedin} > \text{systran} > \text{systran-nrc} > \text{nrc} > \text{saar}
  \]

- **English → French Europarl:**
  
  \[
  \text{LIMSI} > \text{systran-nrc} = \text{uedin} > \text{upc} > \text{nrc} = \text{systran} > \text{saar}
  \]

- **English → French News Corpus:**
  
  \[
  \text{SYSTRAN-NRC=SYSTRAN} > \text{limsi} > \text{nrc} = \text{ucb} = \text{uedin} > \text{ucb} > \text{saar}
  \]
Best Czech-English Systems

- Czech $\rightarrow$ English News Corpus:
  \[\text{UMD} > \text{cu} > \text{uedin} > \text{pct}\]
- English $\rightarrow$ Czech News Corpus:
  \[\text{PCT} > \text{umd} > \text{uedin}\]