Evaluating Evaluation
Lessons from the WMT 2007 Shared Task

Philipp Koehn
(with a lot of help from Chris Callison-Burch)

11 September 2007
WMT Evaluation Campaign

• Annual event since 2005
  – shared task of an ACL Workshop on Machine Translation (WMT)
  – now part of EuroMatrix project
  – about 10–15 groups participate each year
  – next year: ACL 2008 (Ohio), Marathon meeting in May 2008 (Berlin)

• Goals
  – promote MT performance for European languages
  – large-scale (30 million word training), relatively wide domain (politics)
  – low barrier of entry: baseline system provided (Moses)
  – well-defined set of homogenous training data (opposed to NIST eval)
  – allows to for focus on specific problems (e.g. morphology, unknown words)
  – also an opportunity to improve evaluation of MT
Participants

- Some big players missing... (Google, IBM, RWTH Aachen, USC/ISI)
- ... but **wide variety** of systems
  - statistical phrase-based (most)
  - statistical tree-based (CMU)
  - dependency treelet system (Microsoft)
  - rule-based (Systran)
  - rule-based + statistical post-editing (Systran+NRC/Edinburgh)
  - hybrid (system combination) (Saarbrücken)
- Not a toy task: relatively **high translation quality**
Evaluation of Evaluation

- Organizer’s effort mostly focused on questions of evaluation
- **Manual** evaluation: participants volunteer 8 hours worth of work
  - what metric?
  - how many judgments do we need?
  - are judges consistent?
- **Automatic** evaluation
  - evaluation of evaluation campaign
  - what automatic metrics correlate best with human judgment?
WMT Evaluation 2007

- Tasks
  - 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
  - participants in evaluation
  - students of a course on MT
  - paid students at U Edinburgh
Evaluation of Evaluation

• 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?
Fluency and Adequacy

- Traditional metric (used by NIST eval, WMT 2006, IWSLT)
- Two 5-point scales:

<table>
<thead>
<tr>
<th>How much of the meaning of the reference is preserved?</th>
<th>How do you judge the fluency of the translation?</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 = All</td>
<td>5 = Flawless English</td>
</tr>
<tr>
<td>4 = Most</td>
<td>4 = Good English</td>
</tr>
<tr>
<td>3 = Much</td>
<td>3 = Non-native English</td>
</tr>
<tr>
<td>2 = Little</td>
<td>2 = Disfluent English</td>
</tr>
<tr>
<td>1 = None</td>
<td>1 = Incomprehensible</td>
</tr>
</tbody>
</table>
Web Tool

Judge Sentence

You have already judged 14 of 3054 sentences, taking 86.4 seconds per sentence.

**Sources:** 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 laboratory 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

Instructions:
- 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 = Dull/kent English
- 1 = Incomprehensible
Example

In  Le deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l’ue.
Ref  Rather, the two countries form a laboratory needed for the internal working of the EU.

MT 1  Both countries are rather a necessary laboratory the internal operation of the EU.
MT 2  Both countries are a necessary laboratory at internal functioning of the EU.
MT 3  The two countries are rather a laboratory necessary for the internal workings of the EU.
MT 4  The two countries are rather a laboratory for the internal workings of the EU.
MT 5  The two countries are rather a necessary laboratory internal workings of the EU.

Judge each sentence in terms of adequacy and fluency on the scale of 1–5!
### Judgments

<table>
<thead>
<tr>
<th></th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>System 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

- Metrics evaluated by
  - 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 word-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
| people | Iraq | to provide basic services | cannot sustain it | occupation | its | the US | Can
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Können die USA ihre Besetzung aufrechterhalten wenn sie dem irakischen Volk nicht grundlegende Dienstleistungen anbieten können
people's Iraq to services basic provide it sustain its occupation its US the Can

können anbieten Dienstleistungen grundlegende Volk nicht aufrechterhalten Besetzung ihre USA die irakischen sie dem wenn

Philipp Koehn Evaluating Evaluation 11 September 2007
people's services cannot be provided if occupation of the US cannot sustain it.

Können die USA ihre Besetzung aufrechterhalten, wenn sie dem irakischen Volk nicht grundlegende Dienstleistungen anbieten können?
people’s Iraq to services basic provide cannot it if occupation its sustain US the Can the US sustain its occupation if it cannot provide basic services to Iraq ’s people
| Can the US sustain its occupation if it cannot provide basic services to Iraq people's Iraq to services basic provide cannot it if occupation its sustain US the US can maintain its occupation if they cannot offer the Iraqi people basic services |


Philipp Koehn
Evaluating Evaluation
11 September 2007
The University of Edinburgh

Philipp Koehn Evaluating Evaluation 11 September 2007

Können die USA ihre Besetzung aufrechterhalten, wenn sie dem irakischen Volk nicht grundlegende Dienstleistungen anbieten können.

Can the United States maintain its employment, when it cannot provide basic services to people in Iraq?
Results of the Meta-Evaluation

- We measured agreement among annotators using the **kappa coefficient**: 

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

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 \( \kappa \) scores varies, but:
  - .6 – .8 is **good** agreement
  - .4 – .6 is **moderate** agreement
  - < .4 and we should start to **worry**
# Inter-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>.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>

Philipp Koehn
Evaluating Evaluation
11 September 2007
# 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>0.630</td>
<td>0.2</td>
<td>0.537</td>
</tr>
<tr>
<td>Adequacy (absolute)</td>
<td>0.574</td>
<td>0.2</td>
<td>0.468</td>
</tr>
<tr>
<td>Fluency (relative)</td>
<td>0.690</td>
<td>0.333</td>
<td>0.535</td>
</tr>
<tr>
<td>Adequacy (relative)</td>
<td>0.696</td>
<td>0.333</td>
<td>0.544</td>
</tr>
<tr>
<td>Sentence ranking</td>
<td>0.749</td>
<td>0.333</td>
<td>0.623</td>
</tr>
<tr>
<td>Constituent ranking</td>
<td>0.842</td>
<td>0.333</td>
<td>0.762</td>
</tr>
</tbody>
</table>
Time to judge one item

 Constituent rank
 Sentence rank
 Fluency+adequacy scoring

 Num items taking this long (%)

 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>System</th>
<th>Proportion</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%
Carnegie Mellon University 8%
Charles University 4%
University of California at Berkeley 3%
National Research Council + SYSTRAN 2%
SYSTRAN 2%
Saarland University 0.8%
**Systran puzzle (WMT 2006)**

- *English–French, adequacy vs. BLEU, in-domain*

- see also Callison-Burch et al.’s *critique of BLEU* [EACL 2006]
Mystery resolved?

- English–French, adequacy vs. BLEU
- out-of-domain
- Systran: best BLEU, best manual
  → lack of correlation only due to the overly literalness of BLEU?
## 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>0.77</td>
<td>0.84</td>
<td>0.80</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td>ParaEval-Recall</td>
<td>0.71</td>
<td>0.74</td>
<td>0.77</td>
<td>0.80</td>
<td>0.76</td>
</tr>
<tr>
<td>Meteor</td>
<td>0.71</td>
<td>0.72</td>
<td>0.75</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Bleu</strong></td>
<td>0.69</td>
<td>0.72</td>
<td>0.67</td>
<td>0.60</td>
<td>0.67</td>
</tr>
<tr>
<td>Max adeq corr</td>
<td>0.65</td>
<td>0.66</td>
<td>0.66</td>
<td>0.53</td>
<td>0.63</td>
</tr>
<tr>
<td>Max flu corr</td>
<td>0.64</td>
<td>0.65</td>
<td>0.66</td>
<td>0.51</td>
<td>0.61</td>
</tr>
<tr>
<td>GTM</td>
<td>0.66</td>
<td>0.67</td>
<td>0.62</td>
<td>0.50</td>
<td>0.61</td>
</tr>
<tr>
<td>Dependency overlap</td>
<td>0.64</td>
<td>0.64</td>
<td>0.60</td>
<td>0.51</td>
<td>0.60</td>
</tr>
<tr>
<td>ParaEval-Precision</td>
<td>0.64</td>
<td>0.65</td>
<td>0.61</td>
<td>0.49</td>
<td>0.60</td>
</tr>
<tr>
<td>1-TER</td>
<td>0.61</td>
<td>0.54</td>
<td>0.52</td>
<td>0.51</td>
<td>0.54</td>
</tr>
<tr>
<td>1-WER of verbs</td>
<td>0.38</td>
<td>0.42</td>
<td>0.43</td>
<td>0.30</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Semantic Role Overlap

- Proposed by Giménez and Màrquez2007 [WMT 2007]
- Solves the Linear-B [NIST 2005] and Systran [WMT 2006] puzzle
  - NIST 2005: correlation of 0.6–0.7 vs. 0.06 for BLEU
  - WMT 2006: correlation of 0.9–0.95 vs. 0.6–0.85 for BLEU
- Checks if arguments/adjuncts to verbs overlap
- Tunable?
Lessons for Automatic Metrics

- Still an **essential tool** when building SMT systems
- Research papers should also report manual evaluation
- Consistent **bias** in automatic metrics when comparing different type of systems
- Improving automatic evaluation is a **well-defined task**
  - goal: better correlation with human judgments
  - impossible to *game* the metric
  - fast to compute to be usable in tuning
Lessons for Manual Metrics

• Agreement was low for fluency and adequacy scores
• We should research ways of **improving manual evaluation** so that it is
  – more consistent
  – faster / cheaper
  – easier to perform
  – re-usable
• Are we asking the right question?
  – we do **not** care, how **good** machine translation is
  – we do care, how **useful** machine translation is
Future Evaluations

• Euromatrix project starts an **ongoing online evaluation** later this year

• 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 253 language pairs, and 506 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
  – next year will include Hungarian
  – **ideas for manual evaluation welcome!**
Best German-English Systems

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

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

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

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

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

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

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

- English → Spanish News Corpus:
  \[
  \text{SYSTRAN} > \text{upc} > \text{cmu-uka} > \text{ucb} > \text{uedin} > \text{nrc} = \text{upv}
  \]
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:
  \[ \text{SYSTRAN-NRC} = \text{SYSTRAN} > \text{limsi} > \text{nrc} = \text{ucb} = \text{uedin} > \text{ucb} > \text{saar} \]
Best Czech-English Systems

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