Further Meta-Evaluation of Machine Translation

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

19 June 2008
Goals of Shared Task

- Forum for European language translation
  - provide incentive to work on European language translation
  - low barrier of entry → attract one-grad-student teams
  - showcase state of the art

- Openness
  - open to non-statistical systems
  - full release of submissions, results, judgments

- Innovations
  - focus on human evaluation judgments
  - volunteer judging
  - evaluation of evaluation metrics
# Participants

<table>
<thead>
<tr>
<th>BBN, system combination</th>
<th>RBMT: Babelfish, Lingenio, Lucy, OpenLogos, ProMT, SDL (ordering anonymized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carnegie Mellon University, sys. comb.</td>
<td>Systran</td>
</tr>
<tr>
<td>Carnegie Mellon University, Gimpel</td>
<td>Universitat Politecnica de Catalunya, Barcelona</td>
</tr>
<tr>
<td>Carnegie Mellon University, SMT</td>
<td>University of California at Berkeley</td>
</tr>
<tr>
<td>Carnegie Mellon University, Stat-XFER</td>
<td>University College London</td>
</tr>
<tr>
<td>Charles University, TectoMT</td>
<td>University of Edinburgh</td>
</tr>
<tr>
<td>Charles University, Bojar</td>
<td>University of Edinburgh, sys. comb.</td>
</tr>
<tr>
<td>Cambridge University</td>
<td>University of Maryland</td>
</tr>
<tr>
<td>Dublin City University</td>
<td>University of Saarbrücken</td>
</tr>
<tr>
<td>LIMSI, University of Paris</td>
<td>University of Washington</td>
</tr>
<tr>
<td>Linköping University</td>
<td>Xerox Research Centre Europe</td>
</tr>
<tr>
<td>LIUM / Systran</td>
<td></td>
</tr>
</tbody>
</table>
New This Year

- Conditions
  - News test set
  - Hungarian → English, German → Spanish

- Entrants
  - contrastive RBMT runs
  - unofficial system combination entries

- Evaluation
  - official metrics evaluation
  - Yes/No judgments of constituents
  - All → English judgments
Test Sets

- Traditionally: test set taken from reserved part of training data
  - Europarl test set
  - News Commentary test set
- Better: a more representative selection of typical translation task
  - for instance: news stories in European media
  - but no existing translations available
    (even multilingual BBC and EuroNews does not translate)
- EuroMatrix funding allowed creation of news test set
  - stories from six different languages, 300+ sentences each
  - each translated into the other five languages (30 language pairs)
  - co-ordinated by CELCT, with contributions from all EuroMatrix partners
  - total cost: 17,200 euro
Sources for News Test Set

Hungarian: Napi (3 documents), Index (2), Origo (5), Népszabadság (2), HVG (2), Uniospez (1)

Czech: Aktuálně (1), iHNed (4), Lidovky (7), Novinky (3)

French: Liberation (4), Le Figaro (4), Dernieres Nouvelles (2), Les Echos (3), Canoe (2)

Spanish: Cinco Dias (7), ABC.es (3), El Mundo (5)

English: BBC (3), Scotsman (3), Economist (3), Times (3), New York Times (3)

German: Financial Times Deutschland (3), Süddeutsche Zeitung (3), Welt (3), Frankfurter Allgemeine Zeitung (3), Spiegel (3)
System Combination

- Combing the output of all submissions
  - train on Europarl test set entries, translate News test set
  - BBN, CMU, and University of Edinburgh
  - also: submission by University of Saarbrücken

- Results
  - better on BLEU, most other automatic metrics
  - human judgment inconclusive
Shared Task Evaluation

- 156 system submissions were manually evaluated
- Recruited 100+ judges, who contributed 260+ hours for 75,000+ judgments
- Scoring
  - users were asked to rank the same sentence output from up to 5 systems
  - convert to pairwise judgment: how many times was A ranked ≥ B?
  - overall score: how often was A ranked ≥ any other system?
**Source:** Les rues donnant accès à la Cour suprême étaient bloquées lundi matin pour empêcher le rassemblement.

**Reference:** The roads leading to the Supreme Court were blocked on Monday morning for avoiding the meeting.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>The streets giving access to the Supreme Court were blocked on Monday morning to prevent the Assembly.</td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>The streets with access to the Supreme Court were blocked on Monday morning to prevent the gathering.</td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>The streets with access to the Supreme Court were blocked on Monday morning to prevent the rally.</td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>The streets leading to the Supreme Court were blocked Monday morning to prevent the Union.</td>
<td><img src="image" alt="Image" /></td>
</tr>
<tr>
<td>The streets giving access to the High Court were blocked Monday morning to prevent the union.</td>
<td><img src="image" alt="Image" /></td>
</tr>
</tbody>
</table>

**Annotator:** josh  **Task:** WMT08 French-English News

Instructions:
Rank each whole sentence translation from Best to Worst relative to the other choices (ties are allowed).
Results

• Winning systems
  – according to human judgments (sentence ranking)
  – in domain (Europarl, News Commentary)
  – out of domain (News)
• Rule-Based vs. Statistical MT
• What languages are hard to translate?
• For details: see 37 page overview paper
• For more details: look at system outputs and human judgments
## In Domain — Europarl

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Best System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech–English*</td>
<td>Dublin City University</td>
</tr>
<tr>
<td>English-Czech*</td>
<td>Charles University, Bojar</td>
</tr>
<tr>
<td>French-English</td>
<td>LIUM / Systran</td>
</tr>
<tr>
<td>English-French</td>
<td>LIMSI, University of Paris</td>
</tr>
<tr>
<td>German-English</td>
<td>University of Edinburgh</td>
</tr>
<tr>
<td>English-German</td>
<td>RBMT4</td>
</tr>
<tr>
<td>Spanish-English</td>
<td>LIMSI, University of Paris</td>
</tr>
<tr>
<td>English-Spanish</td>
<td>LIMSI, University of Paris</td>
</tr>
</tbody>
</table>

* News Commentary
# Out of Domain — News

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Best System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech–English</td>
<td>Dublin City University</td>
</tr>
<tr>
<td>English-Czech</td>
<td>PC-Translator</td>
</tr>
<tr>
<td>French-English</td>
<td>LIUM / Systran</td>
</tr>
<tr>
<td>English-French</td>
<td>RBMT4</td>
</tr>
<tr>
<td>German-English</td>
<td>RBMT2</td>
</tr>
<tr>
<td>English-German</td>
<td>RBMT2</td>
</tr>
<tr>
<td>Hungarian-English</td>
<td>Morphologic</td>
</tr>
<tr>
<td>Spanish-English</td>
<td>RBMT4</td>
</tr>
<tr>
<td>English-Spanish</td>
<td>RBMT5</td>
</tr>
</tbody>
</table>
## Rule-based vs. Statistical

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>In Domain</th>
<th>Out of Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech-English</td>
<td>SMT</td>
<td>SMT*</td>
</tr>
<tr>
<td>English-Czech</td>
<td>SMT</td>
<td>RBMT</td>
</tr>
<tr>
<td>French-English</td>
<td>SMT</td>
<td>SMT</td>
</tr>
<tr>
<td>English-French</td>
<td>SMT</td>
<td>RBMT</td>
</tr>
<tr>
<td>German-English</td>
<td>SMT</td>
<td>RBMT</td>
</tr>
<tr>
<td>English-German</td>
<td>RBMT</td>
<td>RBMT</td>
</tr>
<tr>
<td>Hungarian-English</td>
<td>-</td>
<td>RBMT</td>
</tr>
<tr>
<td>Spanish-English</td>
<td>SMT</td>
<td>RBMT</td>
</tr>
<tr>
<td>English-Spanish</td>
<td>SMT</td>
<td>RBMT</td>
</tr>
</tbody>
</table>

* no RBMT system in competition
Difficulty of x–English Translation

• Which languages are easiest to translate?
• Languages, sorted by best performing systems for each:

<table>
<thead>
<tr>
<th>Source Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
</tr>
<tr>
<td>Spanish</td>
</tr>
<tr>
<td>German</td>
</tr>
<tr>
<td>Czech</td>
</tr>
<tr>
<td>Hungarian</td>
</tr>
</tbody>
</table>
The “Meta” Part

- We want 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?
Automatic evaluation metrics

- Ranked system outputs using a number of different automatic metrics
  - N-gram matching:
    - *Bleu, Meteor, M-TER, M-Bleu*
  - Linguistic info:
    - *Dependency overlap, Semantic role overlap, Bleu over POS tags*
  - Aggregate measures:
    - *Combination of many of the above*
Evaluating the evaluation metrics

- How well do the automatic metrics correspond to human judgments?
- Compare the metric scores and human rankings (average number of times a system was judged to be better than or equal to any other system)
- Measure Spearman's rank correlation
## Correlation with Sentence Rank

<table>
<thead>
<tr>
<th></th>
<th>DP</th>
<th>ULCh</th>
<th>Meteor-Rank</th>
<th>POS-Bleu</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-English News</td>
<td>.83</td>
<td>.87</td>
<td>.79</td>
<td>.80</td>
<td>.71</td>
</tr>
<tr>
<td>French-English News</td>
<td>.92</td>
<td>.90</td>
<td>.80</td>
<td>.90</td>
<td>.81</td>
</tr>
<tr>
<td>French-English Europarl</td>
<td>.94</td>
<td>.93</td>
<td>.87</td>
<td>.91</td>
<td>.92</td>
</tr>
<tr>
<td>German-English News</td>
<td>.85</td>
<td>.95</td>
<td>.89</td>
<td>.58</td>
<td>.12</td>
</tr>
<tr>
<td>German-English Europarl</td>
<td>.76</td>
<td>.60</td>
<td>.66</td>
<td>.64</td>
<td>.30</td>
</tr>
<tr>
<td>Spanish-English News</td>
<td>.75</td>
<td>.81</td>
<td>.63</td>
<td>.66</td>
<td>.19</td>
</tr>
<tr>
<td>Spanish-English Europarl</td>
<td>.78</td>
<td>.77</td>
<td>.84</td>
<td>.80</td>
<td>.78</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>.83</strong></td>
<td><strong>.83</strong></td>
<td><strong>.78</strong></td>
<td><strong>.76</strong></td>
<td><strong>.55</strong></td>
</tr>
</tbody>
</table>
## Correlation with Sentence Rank

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>POS-Bleu</th>
<th>POSF4Gram-GM</th>
<th>BLEU</th>
<th>SVM-Rank</th>
<th>Meteor-Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-French News</td>
<td>.66</td>
<td>.69</td>
<td><strong>.73</strong></td>
<td>.72</td>
<td>.30</td>
</tr>
<tr>
<td>English-French Europarl</td>
<td>.93</td>
<td><strong>.97</strong></td>
<td>.95</td>
<td>.93</td>
<td>.95</td>
</tr>
<tr>
<td>English-German News</td>
<td><strong>.90</strong></td>
<td><strong>.90</strong></td>
<td>.58</td>
<td>.32</td>
<td>.43</td>
</tr>
<tr>
<td>English-German Europarl</td>
<td>.50</td>
<td>.38</td>
<td>.36</td>
<td><strong>.56</strong></td>
<td>.42</td>
</tr>
<tr>
<td>English-Spanish Europarl</td>
<td>.76</td>
<td>.79</td>
<td>.80</td>
<td>.80</td>
<td><strong>.84</strong></td>
</tr>
<tr>
<td>Average</td>
<td><strong>.75</strong></td>
<td><strong>.75</strong></td>
<td>.68</td>
<td>.67</td>
<td>.59</td>
</tr>
</tbody>
</table>
Examining Manual Evaluation

- Three different types of evaluation:
  - Rank translations of sentences relative to each other
  - Rank translations of sub-sentential units
  - Yes/no judgments on the translations of phrases

- Meta-evaluation measured:
  - Inter-annotator agreement (agreement with others)
  - Intra-annotator agreement (self consistency)
  - Average time to make one judgement
Constituent-based Evaluation

• 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, or give yes/no judgment
people
's
Iraq
to
services
provide
basic
it
cannot
it
occupation
its
sustain
US
Can

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

Können
aufrechterhalten
wenn
sie
dem
irakischen
Volk
nicht
grundlegende
Dienstleistungen
anbieten
können
Can the US sustain its occupation if it cannot provide basic services to Iraq's people?
people's Iraq to provide basic services cannot it sustain its occupation it the US Can

Können die USA ihre Besetzung aufrechterhalten wenn sie dem irakischen Volk nicht grundlegende Dienstleistungen anbieten können.
people's basic services cannot be provided if 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's people?

The US can maintain its occupation if they cannot offer the Iraqi people basic services.
The US occupation cannot maintain its sustain in Iraq. If the people of basic services cannot provide it to the Iraqi, it is clear that the US can maintain its Besetzung aufrechterhalten in Iraq. While the Besetzung aufrechterhalten cannot be maintained, the Volk is not able to sustain grundlegende Dienstleistungen anbieten.
Können die US ihre Besetzung aufrechterhalten, wenn sie dem irakischen Volk nicht grundlegende Dienstleistungen anbieten können?
The United States can maintain its employment when it cannot provide basic services to Iraq's 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:
  – 0.6 – 0.8 is good agreement
  – 0.4 – 0.6 is moderate agreement
  – 0.2 – 0.4 is fair agreement
  – < 0.2 is slight agreement
## 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*</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>.578</td>
<td>.333</td>
<td>.367</td>
</tr>
<tr>
<td>Constituent ranking</td>
<td>.678</td>
<td>.333</td>
<td>.517</td>
</tr>
<tr>
<td>Constituent yes/no</td>
<td>.825</td>
<td>.5</td>
<td>.649</td>
</tr>
</tbody>
</table>

* From ACL-WMT07
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*</td>
<td>.630</td>
<td>.2</td>
<td>.537</td>
</tr>
<tr>
<td>Adequacy*</td>
<td>.574</td>
<td>.2</td>
<td>.468</td>
</tr>
<tr>
<td>Sentence ranking</td>
<td>.691</td>
<td>.333</td>
<td>.537</td>
</tr>
<tr>
<td>Constituent ranking</td>
<td>.832</td>
<td>.333</td>
<td>.748</td>
</tr>
<tr>
<td>Constituent yes/no</td>
<td>.930</td>
<td>.5</td>
<td>.861</td>
</tr>
</tbody>
</table>

* From ACL-WMT07
Time to judge one item

num items taking this long (%) vs. time to judge one item (seconds)

- yes/no judgments
- constituent rank
- sentence rank

Mallison et al. further meta-evaluation of MT
Conclusion

• Surprising results for automatic metrics
  – No single metric always correlates best with human judgments of translation quality

• Making human evaluation easier
  – We can improve both agreement and speed of judgments
  – Is there a downside to focusing on relative ranking?

• Judgments will be downloadable http://www.statmt.org/wmt08/
  – Further analysis
  – Creating new metrics
  – Tuning systems with human judgments
Plans for WMT 09

- Testing period in winter 2008/2009
- Workshop at EACL Athens (March/April 2009)
- Official system combination track
- Reduce tasks to improve statistical significance
  - only news?
- Ongoing online evaluation
  http://matrix.statmt.org/
Discussion

How should we manage the system combination task?
Other human judgment methods?
How can we better deal with sentence-level metrics?
Feedback on the evaluation campaign?