Automatic Evaluation in MT system production

Gr. Thurmair,
Linguatec

MT Summit XI Copenhagen
2007-09-10
Outline

- Quality lifecycle in MT system production
- Automatic metrics methodology
- Example case
Quality Management in MT production

• Automatic evaluation is seen in the context of MT system development
  – linguistic components development

• system development follows general software development technology
  – best ratio between investment and quality improvement
  – planning requirements (time, resources, functionality)
    also for linguistic components
Lingware development cycle

- **Requirement** phase
  - where are significant quality problems in the current system
    - previous evaluations
    - customer feedback
    - competitive evaluations
  - definition of **thematic areas** to work on
    - dictionary coverage
    - long complex sentences
    - 1:n transfers
    - proper name recognition
    - anaphora
    - spelling errors and correction possibilities
Lingware development cycle

• **Specification** phase
  – creation of large **data collections** to study the phenomenon
    • monolingual, bilingual
  
  – Try to find rules / heuristics to 'solve' the problem
    • go through a lot of material (e.g. proper names)
      – not for mark-up for automatic learning
      – but for knowledge extraction

  – specify how the problem can be solved
    • changes / adaptations in dictionaries
    • changes in grammars
    • adding new system components
    • (interaction with other system components)
Lingware development cycle

- **Implementation Phase**
  - create all linguistic resources
    - dictionaries, grammars
    - corpora, training data

- **Component test**
  - create **test material**
    - phenomena in isolation
    - phenomena in context
  - thematic corpora for certain phenomena
  - canonical analysis results or test translations
  - do quality evaluation
    - how many of the analysed phenomena are correct?
    - to which extent can the problem be mastered?
Lingware development cycle

• **System test**
  – lingware
    • side-effects of analysis on other lingware parts?
    • improvement – deterioration analysis
    • overall quality gain
  – overall system
    • interaction with other components
      – dictionary coding, translation memories, ...
    • effect on system performance / resources

• **Evaluation**
  – measure quality improvement
  – start next development cycle
Test corpus

- **Test corpus design**
  - representative for system use
    - **multi-domain**, multi-texttype, multi-purpose
  - fair coverage of linguistic problems
    - sentence length, input errors, `<multi-sentence>`
  - significant size ...
  - all system aspects
    - translation options, additional dictionaries and memories

- **Test corpus creation**
  - reference translations *created by machine*
  - postedited into grammatical sentence
    - faster than human 😊
    - closer to intended purpose
    - (not even required for relative evaluation)
Evaluation

• Quality evaluation
  – *relative* quality („12% better than previous version“)
    • compare with previous system runs
    • 3-point scale (‘better – worse – same’)
      (not every difference deteriorates)
    • quality = % improvement minus % deteriorations
      – depends on type of corpus
  – *absolute* quality („80% quality“)
    • compare with canonical output
    • quality = distance to canonical output
    • quality = related to FEMTI criteria adequacy, fluency
      – 3-point scale (good – understandable – bad)
  – absolute quality percentage does not say too much
    • „we have a translation quality of 68%“: ??
    • depends mainly on test corpus ...
2. Automatic measures

- Where would automatic evaluation be useful?
- Test methodology:
  - create system from training data
    • current setups: what is available
    • MT industry: all customer-relevant domains
  - create test set of reference translations
    • current setups: human reference translators
    • MT industry: MT-produced + post-edited
  - evaluate distance of output to reference
    • current setups: distance measures
    • MT industry: distance, plus: inspection of deviations

"Automatic metrics are not designed to provide direction to R&D" (Miller)
Methodology: reference

- **Reference translations by humans** is problematic
  - "The only professional translator got worse scores than the translations of all seven non-professionals ... This is because the non-professional translations tended to be fairly literal and stayed as close to the source text as possible." (Culey 2003)
  - "The human translations that scored poorly were generally freer translations" (Culey 2003)
  - The better translators are the worse the scores become

- **Number of translations** seems to be less important than closeness
  - (MT-produced output is closest to bad translators 😊)
Methodology: distance

• Pure **word distance** (WER) does not reflect the **quality** of the distance
  – *Take the floppy out of the drive*
    Take the *disquette* out of the drive
    *Take the *elefant* out of the drive*
  – *Hans ging nach Hause zurück*
    John went back home
    John returned to his home
    *John went from home*

• **Relative word order** is meaning-bearing!
  – *the man killed the tiger* - *the tiger killed the man*

• (The score itself does not help much)
4. Example Case

- **Topic:**
  - English–to–Chinese MT system

- **Purpose:**
  - Determine competitive quality of our MT system

- **Evaluated:**
  - three rule-based systems (R1, R2, R3)
  - one statistical system (S1)

(Project Manager: Liu Lezhong)
Test Corpus

- From Chinese Linguistic Data Consortium
  - (ChineseLDC) www.chineseldc.org
  - 2005 863 National Machine Translation Test Set

- English → Chinese
  - 492 sentences with four manual translations

- Chinese → English
  - 489 sentences with four manual translations
Automatic Evaluation

- English -> Chinese

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST</td>
<td>7.1361</td>
<td>8.4120</td>
<td>6.8843</td>
<td>8.2716</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.2426</td>
<td>0.3441</td>
<td>0.2373</td>
<td>0.3699</td>
</tr>
</tbody>
</table>
## Automatic Evaluation

- **Chinese -> English**

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST</td>
<td>5.8890</td>
<td>6.9569</td>
<td>5.5654</td>
<td>7.3221</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.1297</td>
<td>0.1893</td>
<td>0.1210</td>
<td>0.2237</td>
</tr>
</tbody>
</table>
• **Global quality** evaluation
  – Four point scale
    • 1 = syntactically / lexically correct, all information carried over (good)
    • 2 = minor mistakes in lexicon / grammar, most information carried over (understandable)
    • 3 = serious mistakes in lexicon / grammar, little information carried (partly understandable)
    • 4 = rubbish no information carried

• **Best Sentence** Analysis
  – for each sentence: which system produced the best translation

• **Error Analysis**
  – for our candidate: what are the main sources of error?
Human Global Evaluation

- English -> Chinese
Human Global Evaluation

- Chinese -> English
Best Sentence Analysis

- English -> Chinese

![Bar Chart showing comparison of different translation systems or approaches for English-Chinese sentence analysis. The x-axis represents 'The Best Sentence,' and the y-axis shows a range of scores or metrics. The chart includes bars for R1, R2, R3, S1, and None, with R2 having the highest value.]
Best Sentence Analysis

- Chinese -> English
• Dictionary work
  – dictionary gaps -> increase to > 400K
  – wrong transfer selection (1:n translations): neural transfer
Result

- Relation of automatic vs. human scores is not stable
  - some papers: it is parallel, others: it is different

- Could be due to the measure
  - word difference (WER) can be positive or negative
  - strongly depends on the reference translation

- Automatic score does not really help
  - “We have translation quality of 63%”: ??
  - there are too many parameters to be taken into account

- System development needs hints for improvement
  - evaluation and error analysis require human intelligence
Task-based evaluation

• Machine Translation is successful if it achieves **productivity gain**
  – Translation of more material in shorter time

• This is how the market decides

• It is **not a quality measure**!

• Productivity can be increased by many means
  – good editor support
  – translation memories, fast dictionary update, ...
Thank you for your attention

g.thurmair@linguatec.de
Quality criteria: Fluent but inadequate

(Source) 倒塌居民房屋2.6万间，损坏房屋59455
(human) Destroyed houses amounted to 260,000, and damaged ones numbered at 59,455.
(SMT) 26,000 housing residents collapse, damaged houses 59455 companies.

(Source) 心理医生指出，很多人胖是因为能吃。
(human) Psychologists point out that obesity is linked to one's capacity to eat.
(SMT) psychologists noted that because many people did gain.

(Source) 全市的6条省级以上交通干线和近30条普通公路受损，9座桥涵被毁。
(human) Six provincial highways and 30 public roads in the city were destroyed, and nine bridges were destroyed. </s>
(SMT) the six provincial highway and nearly 30% of ordinary roads damaged, 3,250 destroyed nine blocks.
Alternatives

- Can formal measures provide (indirect) indication for quality?
  - conclude overall quality from „easy“ domain
    - named entity translation
    - noun phrases / compounds
  - still we would want to have the kind of errors