Automatic Evaluation of Machine Translation Quality

Lluís Màrquez
TALP Research Center
Technical University of Catalonia (UPC)

Invited talk at Dialogue 2013
Bekasovo Resort, Russia
May 30, 2013
Joint work with:

Jesús Giménez, Lluís Formiga and Meritxell Gonzàlez
1 Automatic MT Evaluation
2 Linguistically-motivated Measures
3 Intelligent MT output and error analysis
4 Quality Estimation
Talk Overview

1. Automatic MT Evaluation
2. Linguistically-motivated Measures
3. Intelligent MT output and error analysis
4. Quality Estimation
MT System Development Cycle

Error Detection

Error Analysis

Refinement

Implementation

Test

OK? Yes

Discard

Unfruitful Results

Keep

Evaluation Methods

MT System Developer
Difficulties of MT Evaluation

- Machine Translation is an open NLP task
  - the *correct translation* is not unique
  - the set of valid translations is not small
  - translation correctness is not black and white

- Quality aspects are *heterogeneous*
  - Adequacy (or Fidelity)
  - Fluency (or Intelligibility)
  - Post-editing effort (time, key strokes, ...)
  - ...

- Manual vs. automatic evaluation
MT Automatic Evaluation

Setting:

⇒ Compute similarity between system’s output and one or several reference translations

⇒ The similarity measure should be able to discriminate whether the two sentences convey the same meaning (semantic equivalence)
Setting:

⇒ Compute similarity between system’s output and one or several reference translations

Challenge:

⇒ The similarity measure should be able to discriminate whether the two sentences convey the same meaning (semantic equivalence)
First Approaches:

⇒ *Lexical similarity* as a measure of quality
MT Automatic Evaluation

First Approaches:

⇒ Lexical similarity as a measure of quality

• Edit Distance
  WER, PER, TER

• Precision
  BLEU, NIST, WNM

• Recall
  ROUGE, CDER

• Precision/Recall
  GTM, METEOR, BLANC, SIA
MT Automatic Evaluation

First Approaches:

⇒ Lexical similarity as a measure of quality

- Edit Distance
  WER, PER, TER

- Precision
  BLEU, NIST, WNM

- Recall
  ROUGE, CDER

- Precision/Recall
  GTM, METEOR, BLANC, SIA

- **BLEU** has been widely accepted as a ‘de facto’ standard
The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family.”
Conclusions of the paper (Papineni et al., 2001)

- BLEU correlates with human judgements
- It can distinguish among similar systems
- Need for multiple references or a big test with heterogeneous references
- More parametrisation in the future
Benefits of Automatic Evaluation

Compared to manual evaluation, automatic measures are:

1. **Cheap** (vs. costly)
2. **Objective** (vs. subjective)
3. **Reusable** (vs. not-reusable)

Automatic evaluation metrics have notably accelerated the development cycle of MT systems

1. Error analysis
2. System optimization
3. System comparison
Benefits of Automatic Evaluation

Compared to manual evaluation, automatic measures are:

1. Cheap (vs. costly)
2. Objective (vs. subjective)
3. Reusable (vs. not-reusable)

Automatic evaluation metrics have notably accelerated the development cycle of MT systems:

1. Error analysis
2. System optimization
3. System comparison
Risks of Automatic Evaluation (compared to manual evaluation)

1. **System overtuning** → when system parameters are adjusted towards a given metric

2. **Blind system development** → when metrics are unable to capture actual system improvements

3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems
Risks of Automatic Evaluation (compared to manual evaluation)

1. **System overtuning** → when system parameters are adjusted towards a given metric

2. **Blind system development** → when metrics are unable to capture actual system improvements

3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems
Risks of Automatic Evaluation (compared to manual evaluation)

1. **System overtuning** → when system parameters are adjusted towards a given metric

2. **Blind system development** → when metrics are unable to capture actual system improvements

3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems
Risks of Automatic Evaluation (compared to manual evaluation)

1. **System overtuning** → when system parameters are adjusted towards a given metric

2. **Blind system development** → when metrics are unable to capture actual system improvements

3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems
Lexical similarity is not a sufficient nor a necessary condition so that two sentences express the same meaning (Culy and Riehemann, 2003; Coughlin, 2003; Callison-Burch et al., 2006).

The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations.

Lexical metrics have problems distinguishing MT output from fully fluent and adequate translations obtained from them through professional postediting (Denkowski and Lavie, 2012).
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)

⇒ $n$-gram based metrics favor MT systems which closely replicate the lexical realization of the references

⇒ Test sets tend to be similar (domain, register, sublanguage) to training materials

⇒ Statistical MT systems heavily rely on the training data

⇒ Statistical MT systems tend to share the reference sublanguage and be favored by $n$-gram based measures
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)

⇒ *n*-gram based metrics favor MT systems which closely replicate the lexical realization of the references

⇒ Test sets tend to be similar (domain, register, sublanguage) to training materials

⇒ Statistical MT systems heavily rely on the training data

⇒ Statistical MT systems tend to share the reference sublanguage and be favored by *n*-gram based measures
Talk Overview

1. Automatic MT Evaluation
2. Linguistically-motivated Measures
3. Intelligent MT output and error analysis
4. Quality Estimation
Can we do better?

1. Compare to a very large set of references

   - HyTER \((\text{Dreyer and Marcu, 2012})\)
     
     - Construct for every test case a compact network encoding an exponentially large number of meaning equivalent reference translations
     
     - Compute a TER-based similarity over the whole set of translation equivalents
     
     - HyTER correlates much better with human assessments
     
     - But the cost of generating the graphs is very high
Can we do better?

1. Compare to a very large set of references

   - **HyTER (Dreyer and Marcu, 2012)**
     
     ⇒ Construct for every test case a compact network encoding an exponentially large number of meaning equivalent reference translations
     
     ⇒ Compute a TER-based similarity over the whole set of translation equivalents
     
     ⇒ HyTER correlates much better with human assessments
     
     ⇒ But the cost of generating the graphs is very high
2. Generalize over lexical matching

- Lexical variants
  - Morphological information (i.e., stemming)
    - ROUGE and METEOR
  - Synonymy lookup: METEOR (based on WordNet)

- Paraphrasing support:
  - (Zhou et al., 2006; Kauchak and Barzilay, 2006; Owczarzak et al., 2006)
  - Recent versions of METEOR, TER
3. **More linguistically-motivated measures**

- Features capturing *syntactic* and *semantic* information

- Shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, discourse representation

- Very extensive bibliography in the last years
  Check ([Giménez and Màrquez 2010](#)) for a survey
Some Examples of Linguistically Motivated Measures

- **Expected Dependency Pair Match**  
  (Kahn, Snover and Ostendorf, 2009)
  - dependency parsing (PCFG + head-finding rules)
  - precision and recall scores of various tree decompositions
  - +synonymy +paraphrasing

- **MaxSim**  (Chen and Ng; 2008)
  - a general framework for arbitrary similarity functions
  - dependency relations, lemma, parts of speech, synonymy
  - bipartite graph to obtain an optimal matching between items

- **RTE**  (Padó, Galley, Jurafsky and Manning, 2009)
  - semantic equivalence based on textual entailment features
  - alignment, semantic compatibility, insertion/deletion, preservation of reference and structural alignment
Some Examples of Linguistically Motivated Measures

- **Expected Dependency Pair Match**
  (Kahn, Snover and Ostendorf, 2009)
  - dependency parsing (PCFG + head-finding rules)
  - precision and recall scores of various tree decompositions
  - synonymy + paraphrasing

- **MaxSim** (Chen and Ng; 2008)
  - a general framework for arbitrary similarity functions
  - dependency relations, lemma, parts of speech, synonymy
  - bipartite graph to obtain an optimal matching between items

- **RTE** (Padó, Galley, Jurafsky and Manning, 2009)
  - semantic equivalence based on textual entailment features
  - alignment, semantic compatibility, insertion/deletion, preservation of reference and structural alignment
Work at UPC with Jesús Giménez

Rather than comparing sentences at lexical level:

**Compare the linguistic structures and the words within them**
<table>
<thead>
<tr>
<th><strong>Automatic Translation</strong></th>
<th>On Tuesday several missiles and mortar shells fell in south Kabul, but there were no casualties.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference Translation</strong></td>
<td>Several rockets and mortar shells fell today, Tuesday, in south Kabul without causing any casualties.</td>
</tr>
</tbody>
</table>
Our Approach

(Giménez & Màrquez, 2010)

On Tuesday, several missiles and mortar shells fell in south Kabul, but there were no casualties.
Our Approach

(Giménez & Màrquez, 2010)
Measuring Structural Similarity

- **OVERLAP**: generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient

- **Linguistic element** (LE) = abstract reference to any possible type of linguistic unit, structure, or relationship among them
  - For instance: POS tags, word lemmas, NPs, syntactic phrases
  - A sentence can be seen as a bag (or a sequence) of LEs of a certain type
  - LEs may embed
Measuring Structural Similarity

- **OVERLAP**: generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient

- **Linguistic element** (LE) = abstract reference to any possible type of linguistic unit, structure, or relationship among them
  - For instance: POS tags, word lemmas, NPs, syntactic phrases
  - A sentence can be seen as a bag (or a sequence) of LEs of a certain type
  - LEs may embed
Overlap among Linguistic Elements

\[ O(t) = \frac{\sum_{i \in (\text{items}_t(\text{hyp}) \cap \text{items}_t(\text{ref}))} \text{count}_{\text{hyp}}(i, t)}{\sum_{i \in (\text{items}_t(\text{hyp}) \cup \text{items}_t(\text{ref}))} \max(\text{count}_{\text{hyp}}(i, t), \text{count}_{\text{ref}}(i, t))} \]

\( t \) is the LE type

‘hyp’: hypothesized translation

‘ref’: reference translation

\( \text{items}_t(s) \): set of items occurring inside LEs of type \( t \)

\( \text{count}_s(i, t) \): occurrences of item \( i \) in \( s \) inside a LE of type \( t \)
Overlap among Linguistic Elements

Coarser variant: micro-averaged overlap over all types

\[
O(\ast) = \frac{\sum_{t \in T} \sum_{i \in (\text{items}_t(\text{hyp}) \cap \text{items}_t(\text{ref}))} \text{count}_{\text{hyp}}(i, t)}{\sum_{t \in T} \sum_{i \in (\text{items}_t(\text{hyp}) \cup \text{items}_t(\text{ref}))} \max(\text{count}_{\text{hyp}}(i, t), \text{count}_{\text{ref}}(i, t))}
\]

\(T\): set of all LE types associated to the given LE class
Overlap/Matching among Linguistic Elements

- **Matching** is a similar but more strict variant
  - All items inside an element are considered the same unit
  - Computes the proportion of fully translated LEs, according to their types

- Other possible extensions:
  - $n$-gram matching within LEs
  - Synonymy lookup
Matching is a similar but more strict variant

⇒ All items inside an element are considered the same unit
⇒ Computes the proportion of fully translated LEs, according to their types

Other possible extensions:

⇒ n-gram matching within LEs
⇒ Synonymy lookup
Overlap and Matching have been instantiated over different linguistic level elements (for English)

- Words, lemmas, POS
- Shallow, dependency and constituency parsing
- Named entities and semantic roles
- Discourse representation (logical forms)
Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)
### Evaluating Heterogeneous Features

**NIST 2005 Arabic-to-English Exercise**

<table>
<thead>
<tr>
<th>Level</th>
<th>Metric</th>
<th>$\rho_{all}$</th>
<th>$\rho_{SMT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td>BLEU</td>
<td>0.06</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>0.05</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>Syntactic</strong></td>
<td>Parts-of-speech</td>
<td>0.42</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>Dependencies (HWC)</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Constituents (STM)</td>
<td>0.74</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Semantic</strong></td>
<td>Semantic Roles</td>
<td>0.72</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Discourse Repr.</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Discourse Repr. (PoS)</td>
<td>0.97</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

<table>
<thead>
<tr>
<th>Level</th>
<th>Metric</th>
<th>$\rho_{all}$</th>
<th>$\rho_{SMT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td>BLEU</td>
<td>0.06</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>0.05</td>
<td><strong>0.90</strong></td>
</tr>
<tr>
<td><strong>Syntactic</strong></td>
<td>Parts-of-speech</td>
<td>0.42</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>Dependencies (HWC)</td>
<td><strong>0.88</strong></td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Constituents (STM)</td>
<td>0.74</td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td><strong>Semantic</strong></td>
<td>Semantic Roles</td>
<td>0.72</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>Discourse Repr.</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Discourse Repr. (PoS)</td>
<td><strong>0.97</strong></td>
<td>0.90</td>
</tr>
</tbody>
</table>
## Evaluating Heterogeneous Features

### NIST 2005 Arabic-to-English Exercise

<table>
<thead>
<tr>
<th>Level</th>
<th>Metric</th>
<th>$\rho_{\text{all}}$</th>
<th>$\rho_{\text{SMT}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>BLEU</td>
<td>0.06</td>
<td>0.83</td>
</tr>
<tr>
<td>Lexical</td>
<td>METEOR</td>
<td>0.05</td>
<td>0.90</td>
</tr>
<tr>
<td>Syntactic</td>
<td>Parts-of-speech</td>
<td>0.42</td>
<td>0.89</td>
</tr>
<tr>
<td>Syntactic</td>
<td>Dependencies (HWC)</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>Syntactic</td>
<td>Constituents (STM)</td>
<td>0.74</td>
<td>0.95</td>
</tr>
<tr>
<td>Semantic</td>
<td>Semantic Roles</td>
<td>0.72</td>
<td>0.96</td>
</tr>
<tr>
<td>Semantic</td>
<td>Discourse Repr.</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Semantic</td>
<td>Discourse Repr. (PoS)</td>
<td><strong>0.97</strong></td>
<td>0.90</td>
</tr>
</tbody>
</table>
Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

<table>
<thead>
<tr>
<th>Level</th>
<th>Metric</th>
<th>$\rho_{all}$</th>
<th>$\rho_{SMT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td>BLEU</td>
<td>0.06</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>0.05</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>Syntactic</strong></td>
<td>Parts-of-speech</td>
<td>0.42</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>Dependencies (HWC)</td>
<td><strong>0.88</strong></td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Constituents (STM)</td>
<td>0.74</td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td><strong>Semantic</strong></td>
<td>Semantic Roles</td>
<td>0.72</td>
<td><strong>0.96</strong></td>
</tr>
<tr>
<td></td>
<td>Discourse Repr.</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Discourse Repr. (PoS)</td>
<td><strong>0.97</strong></td>
<td>0.90</td>
</tr>
</tbody>
</table>
Towards Heterogeneous Automatic MT Evaluation

Lexical Precision
Lexical Recall
F-measure
Edit Distance
PoS Tagging
Dependency Parsing
Chunking
Constituency Parsing
Lemmatization
Semantic Roles
Named Entities
Discourse Representations

Lexical Similarity
Syntactic Similarity
Semantic Similarity
Towards Heterogeneous Automatic MT Evaluation

Lexical-motivated Measures

Linguistically-motivated Measures

Towards Heterogeneous Automatic MT Evaluation

Lexical Precision

SP-NISTc

SP-Op-*

DP-Or-*

DP-Oc-*

HWCM

Named Entities

NER

BLEU

SIA

METEOR

F-measure

CDER

GTM

BLANC

TER

WER

Lexical Similarity

Syntactic Similarity

Semantic Similarity

DP-OI-*

Dependency Parsing

BLEUATRE

MAXSIM

Semantic Roles

Constituency Parsing

CRP-Or*

DR-STM

DR-Orp-*

Semantic Similarity
Combined Evaluation Measures

- Different measures capture different aspects of similarity
  Suitable for combination

- Extense bibliography on learning to combine evaluation measures. Check (Giménez and Màrquez 2010) for a survey
The Most Simple Approach: ULC

- Uniformly averaged linear combination of measures (ULC):
  \[ \text{ULC}_M(hyp, \text{ref}) = \frac{1}{|M|} \sum_{m \in M} m(hyp, \text{ref}) \]

- Simple hill climbing approach to find the best subset of measures \( M \) on a development corpus

\[ M = \{ \text{‘ROUGE}_W’, \text{‘METEOR’, ‘DP-HWC}_r’, \text{‘DP-O}_c(\star’), \text{‘DP-O}_l(\star’), \text{‘DP-O}_r(\star’), \text{‘CP-STM}_4’, \text{‘SR-O}_r(\star’), \text{‘SR-O}_rv’, \text{‘DR-O}_rp(\star’) \} \]
The Most Simple Approach: ULC

- Uniformly averaged linear combination of measures (ULC):

\[
ULC_M(\text{hyp}, \text{ref}) = \frac{1}{|M|} \sum_{m \in M} m(\text{hyp}, \text{ref})
\]

- Simple hill climbing approach to find the best subset of measures \(M\) on a development corpus

\[M = \{ \text{’ROUGE}_W\text{’, ’METEOR’, ’DP-HWC}_r\text{’, ’DP-O}_c(\star)\text{’, ’DP-O}_l(\star)\text{’, ’DP-O}_r(\star)\text{’, ’CP-STM}_4\text{’, ’SR-O}_r(\star)\text{’, ’SR-O}_rv\text{’, ’DR-O}_rp(\star)\} \]
## Evaluation of ULC

WMT 2008 meta-evaluation results (into-English)

<table>
<thead>
<tr>
<th>Measure</th>
<th>$\rho_{sys}$</th>
<th>consistency$_{snt}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ULC</td>
<td>0.83</td>
<td>0.56</td>
</tr>
<tr>
<td>DP-O$_{r}(\star)$</td>
<td>0.83</td>
<td>0.51</td>
</tr>
<tr>
<td>DR-O$_{r}(\star)$</td>
<td>0.80</td>
<td>0.50</td>
</tr>
<tr>
<td>METEOR$_{ranking}$</td>
<td>0.78</td>
<td>0.51</td>
</tr>
<tr>
<td>SR-O$_{r}(\star)$</td>
<td>0.77</td>
<td>0.50</td>
</tr>
<tr>
<td>METEOR$_{baseline}$</td>
<td>0.75</td>
<td>0.51</td>
</tr>
<tr>
<td>PoS-BLEU</td>
<td>0.75</td>
<td>0.44</td>
</tr>
<tr>
<td>PoS-4gram-F</td>
<td>0.74</td>
<td>0.50</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.52</td>
<td>—</td>
</tr>
<tr>
<td>BLEU$_{stem+wnsyn}$</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Evaluation of ULC

WMT 2009 meta-evaluation results (into-English)

<table>
<thead>
<tr>
<th>Measure</th>
<th>$\rho_{sys}$</th>
<th>$\text{consistency}_{\text{snt}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ULC</strong></td>
<td>0.83</td>
<td>0.54</td>
</tr>
<tr>
<td>maxsim</td>
<td>0.80</td>
<td>0.52</td>
</tr>
<tr>
<td>rte (absolute)</td>
<td>0.79</td>
<td>0.53</td>
</tr>
<tr>
<td>meteor-rank</td>
<td>0.75</td>
<td>0.49</td>
</tr>
<tr>
<td>rte (pairwise)</td>
<td>0.75</td>
<td>0.51</td>
</tr>
<tr>
<td>terp</td>
<td>-0.72</td>
<td>0.50</td>
</tr>
<tr>
<td>meteor-0.6</td>
<td>0.72</td>
<td>0.49</td>
</tr>
<tr>
<td>meteor-0.7</td>
<td>0.66</td>
<td>0.49</td>
</tr>
<tr>
<td>bleu-ter/2</td>
<td>0.58</td>
<td>—</td>
</tr>
<tr>
<td>nist</td>
<td>0.56</td>
<td>—</td>
</tr>
<tr>
<td>wpF</td>
<td>0.56</td>
<td>0.52</td>
</tr>
<tr>
<td>ter</td>
<td>-0.54</td>
<td>0.45</td>
</tr>
</tbody>
</table>

...
Portability Across Corpora

NIST 2004/2005 MT Evaluation Campaigns

<table>
<thead>
<tr>
<th></th>
<th>$\text{AE}_{2004}$</th>
<th>$\text{CE}_{2004}$</th>
<th>$\text{AE}_{2005}$</th>
<th>$\text{CE}_{2005}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#references</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>#outputs$_{\text{ass.}}$</td>
<td>5/5</td>
<td>10/10</td>
<td>6/7</td>
<td>5/10</td>
</tr>
<tr>
<td>#sentences$_{\text{ass.}}$</td>
<td>347/1,353</td>
<td>447/1,788</td>
<td>266/1,056</td>
<td>272/1,082</td>
</tr>
<tr>
<td>Avg. Adequacy</td>
<td>2.81/5</td>
<td>2.60/5</td>
<td>3.00/5</td>
<td>2.58/5</td>
</tr>
<tr>
<td>Avg. Fluency</td>
<td>2.56/5</td>
<td>2.41/5</td>
<td>2.70/5</td>
<td>2.47/5</td>
</tr>
</tbody>
</table>
Portability Across Corpora

Meta-evaluation of ULC across test beds
(Pearson Correlation)

<table>
<thead>
<tr>
<th></th>
<th>AE04</th>
<th>CE04</th>
<th>AE05</th>
<th>CE05</th>
</tr>
</thead>
<tbody>
<tr>
<td>ULC (AE04)</td>
<td>0.6392</td>
<td>0.6294</td>
<td>0.5327</td>
<td>0.5695</td>
</tr>
<tr>
<td>ULC (CE04)</td>
<td>0.6306</td>
<td>0.6333</td>
<td>0.5115</td>
<td>0.5692</td>
</tr>
<tr>
<td>ULC (AE05)</td>
<td>0.6175</td>
<td>0.6029</td>
<td>0.5450</td>
<td>0.5706</td>
</tr>
<tr>
<td>ULC (CE05)</td>
<td>0.6218</td>
<td>0.6208</td>
<td>0.5270</td>
<td>0.6047</td>
</tr>
<tr>
<td>Max Indiv.</td>
<td>0.5877</td>
<td>0.5955</td>
<td>0.4960</td>
<td>0.5348</td>
</tr>
</tbody>
</table>
Many MT evaluation campaigns have been conducted in the last years under NIST, WMT and IWSLT events.

Controversial results at NIST Metrics MATR08/09 Challenges, with bad results in general for linguistic–based evaluation measures.

Finding a practical robust automatic evaluation metric, which correlates well with human assessments is still an open problem.
Summary

1. Evaluation methods play a crucial role

2. Measuring overall translation quality is hard
   ⇒ Quality aspects are heterogeneous and diverse

3. What can we do?
   ⇒ Advance towards heterogeneous evaluation methods
   ⇒ Metricwise system development
      Always meta-evaluate
      (make sure your metric fits your purpose)
   ⇒ Resort to manual evaluation
      Always conduct manual evaluations
      (contrast your automatic evaluations)
      Always do error analysis (semi-automatic)
Talk Overview

1. Automatic MT Evaluation
2. Linguistically-motivated Measures
3. Intelligent MT output and error analysis
4. Quality Estimation
**MT output and error analysis**

**Asiya: An Open Toolkit for Automatic MT Evaluation**

⇒ Integrates all the evaluation measures from (Giménez and Màrquez, 2010)

⇒ **Goal**: to facilitate a practical analysis of large and complex test suites, along several dimensions

  ▶ System evaluation and comparison with a rich family of metrics
  ▶ Error analysis
  ▶ Meta-evaluation of evaluation metrics

⇒ Useful for MT system and evaluation metric developers

⇒ Available and downloadable from:

  [http://www.lsi.upc.es/~nlp/Asiya/](http://www.lsi.upc.es/~nlp/Asiya/)
Recent developments

⇒ **ASIYA in the cloud** (Gonzàlez et al., 2012;2013)

  1. **ASIYA Web Service**
  2. **ASIYA Online Interface**
  3. **ASIYA tSEARCH module**

⇒ Demo video at the same **ASIYA** website
Talk Overview

1. Automatic MT Evaluation
2. Linguistically-motivated Measures
3. Intelligent MT output and error analysis
4. Quality Estimation
Translation Quality Estimation

Quality Estimation (QE)

⇒ Estimate translation quality without reference translations

⇒ Information available
  ▶ Source sentence, candidate translation(s), and some MT system information

⇒ Application scenarios
  ▶ Informing MT end-users about estimated translation quality
  ▶ Quality-oriented filtering of translated texts
    ⇒ identify translations requiring manual post-edition
    ⇒ identify useful post-editions from users
  ▶ Ranking of several translation alternatives
    ⇒ system selection, parameter optimization
Translation Quality Estimation

Quality Estimation (QE)

⇒ Estimate translation quality without reference translations

⇒ Information available
  ▶ Source sentence, candidate translation(s), and some MT system information

⇒ Application scenarios
  ▶ Informing MT end-users about estimated translation quality
  ▶ Quality-oriented filtering of translated texts
    ⇒ identify translations requiring manual post-editing
    ⇒ identify useful post-editions from users
  ▶ Ranking of several translation alternatives
    ⇒ system selection, parameter optimization
Translation Quality Estimation

QE approaches

⇒ Scoring task to predict the absolute quality of the automatic translation of an input text

▷ Usually implemented as a regression function
▷ Also as a direct ranking between translation alternatives
▷ Supervised learning from a training set with human assessments
Translation Quality Estimation

Relevant work

⇒ Johns Hopkins University Summer Workshop, 2003
   “Confidence Estimation for Machine Translation”
   (Blatz et al., 2003)

⇒ Recent work:
   (Specia et al., 2009;2010), (Soricut and Echihabi, 2010),
   (Giménez and Specia 2010), (Pighin et al., 2011),
   (Avramidis, 2012), etc.

⇒ WMT 2012 shared task on Quality Estimation
   (Callison-Burch et al., 2012) (2nd edition at WMT 2013)
Features to train the QE measures

- System–dependent
- System–independent
Features to train the QE measures

- System–dependent
  - internal system probabilities/scores
  - features over $n$-best translation hypotheses
    - language modeling
    - hypothesis rank
    - score ratio
    - average hypothesis length
    - length ratio
    - center hypothesis

- System–independent
Quality Estimation

**Features** to train the QE measures

- **System–dependent**
- **System–independent**
  
  ⇒ **Source** (translation *difficulty*)
  
  ▶ sentence length
  
  ▶ ambiguity → dictionary/alignment/WordNet-based
    (number of candidate translations per word or phrase)
  
  ⇒ **Target** (translation *fluency*)
  
  ▶ sentence length
  
  ▶ language modeling
  
  ⇒ **Source-Target** (translation *adequacy*)
  
  ▶ length ratio
  
  ▶ punctuation issues
  
  ▶ candidate matching → dictionary-/alignment-based
Translation Quality Estimation

QE challenges

⇒ QE is as difficult as MT itself!

⇒ Real adequacy–based QE measures are difficult to apply
  ▶ Training sets are small
  ▶ Involving sophisticated linguistic knowledge easily leads to severe data sparseness
The FAUST Project (2010-2013)

- **Feedback Analysis for User Adaptive Statistical Translation**
- **FP7-ICT-2009-4** (Language-based interaction)
- **http://divf.eng.cam.ac.uk/faust**

**Goal** Develop interactive machine translation systems which adapt rapidly and intelligently to user feedback

- **Challenges in FAUST: real life MT**
  - Open general translation
  - Casual users (feedback is unreliable)
  - Non-standard and noisy translation texts
  - Rapid integration of feedback is required
The FAUST Project (2010-2013)

- Feedback Analysis for User Adaptive Statistical Translation
- FP7-ICT-2009-4 (Language-based interaction)
- http://divf.eng.cam.ac.uk/faust

**Goal** Develop interactive machine translation systems which adapt rapidly and intelligently to user feedback

**Challenges in FAUST: real life MT**
- Open general translation
- Casual users (feedback is unreliable)
- Non-standard and noisy translation texts
- Rapid integration of feedback is required
Learning Quality Estimation measures (FAUST)

**Task**  Training a combination of simple QE features to produce better predictors of translation quality on FAUST data

**Setting**

⇒ We used human feedback in the form of translation quality pairwise rankings. FAUST benchmark corpus: ∼1,900 input segments (en-es), translated by 5 MT systems

⇒ Use of several feature families. Some novel

⇒ Regression vs. ranking SVM learning

⇒ Evaluation in terms of:
  ▶ Correlation of the predicted rankings with the gold standard
  ▶ *Selection of the best translation* (system combination)
Learning Quality Estimation measures (FAUST)

We considered features from 4 different families

1. **Specia Baseline (17) (Specia et al., 2010)**
   - token counts and their ratio, LM probabilities, \( n \)-grams filtered by quartiles, punctuation marks and fertility ratios

2. **ASIYA QE features (26) (González et al., 2012)**
   - bilingual dictionary ambiguity and overlap; overlap ratios on chunks, named-entities and PoS; source and candidate language model perplexities and inverse perplexities over lexical forms, chunks and PoS and out-of-vocabulary word indicators

3. **Features based on adapted Language Models (2)**
   - Words and POS tags. Interpolation weights were computed as to minimize the perplexity according to the Spanish FAUST development set
Learning Quality Estimation measures (FAUST)

We considered features from 4 different families

4. Pseudo-reference based features (Soricut and Echihabi, 2010)

- **Idea**: automatically produced translations by other systems are taken as references

- **Rationale**: if system $X$ produced a translation $A$ and system $Y$ produced a translation $B$ starting from the same input, and $A$ and $B$ are similar and $X$ and $Y$ are different systems, then $A$ is probably a good translation

- Calculated with **BLEU**, **NIST**, **METEOR**, etc. (5) but also with the linguistic-based metrics from **ASIYA** (23)
Main Results on FAUST test data

⇒ It is possible to learn reasonably good QE models from the FAUST annotated corpus, exhibiting fair correlation with the gold-standard rankings

⇒ For the system selection task, pairwise ranking yields better results than regression

⇒ Results are clearly over the baselines. They are also slightly over the system-informed Oracle-Dominant

⇒ All proposed extensions of the basic feature set were useful to boost the quality of the QE modelssystem selection task
Learning Quality Estimation measures (FAUST)

- Quality of the predicted rankings
  - Spearman correlation ($\rho$): 33.86 – 38.43
  - Kendall correlation ($\tau$): 29.67 – 33.02
  - Accuracy of pairwise rankings: 44.67 – 58.11
  - Accuracy at predicting best translation: 39.44 – 51.11

- Results on the system selection task
Learning Quality Estimation measures (FAUST)

- Quality of the predicted rankings
  - Spearman correlation ($\rho$): 33.86 – **38.43**
  - Kendall correlation ($\tau$): 29.67 – **33.02**
  - Accuracy of pairwise rankings: 44.67 – 58.11
  - Accuracy at predicting best translation: 39.44 – **51.11**

- Results on the system selection task

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Ranker</th>
<th>OracleD</th>
<th>OracleB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bleu</strong></td>
<td>33.64</td>
<td><strong>38.28</strong></td>
<td>37.57</td>
<td>44.91</td>
</tr>
<tr>
<td><strong>Meteor</strong></td>
<td>48.34</td>
<td><strong>54.19</strong></td>
<td>54.09</td>
<td>58.15</td>
</tr>
<tr>
<td><strong>Nist</strong></td>
<td>33.64</td>
<td><strong>38.28</strong></td>
<td>37.57</td>
<td>44.91</td>
</tr>
</tbody>
</table>
Learning Quality Estimation measures (FAUST)

Contribution of every family of features
Thank you!

Automatic Evaluation of Machine Translation Quality

Lluís Màrquez
TALP Research Center
Technical University of Catalonia (UPC)

Invited talk at Dialogue 2013
Bekasovo Resort, Russia
May 30, 2013