Automatic Evaluation in Machine Translation
Towards Combined Linguistically-motivated Measures

Lluís Màrquez and Jesús Giménez
TALP Research Center
Tecnhical University of Catalonia

Machine Translation and Morphologically-rich Languages
Research Workshop of the Israel Science Foundation
University of Haifa, January 24, 2010
1 Automatic MT Evaluation
2 Combined Linguistically-motivated Measures
3 Confidence Estimation
4 Conclusions
Talk Overview

1. Automatic MT Evaluation
2. Combined Linguistically-motivated Measures
3. Confidence Estimation
4. Conclusions
MT System Development Cycle

1. Error Detection
2. Error Analysis
3. Refinement
4. Implementation
5. Test

- Keep (YES)
- Discard (NO)

Evaluation Methods

Unfruitful Results
Machine Translation is an *open* NLP task

- the *correct translation* is not unique
- the set of valid translations is not small
- the *quality* of a translation is a fuzzy concept

Quality aspects are *heterogeneous*

- Adequacy (or Fidelity)
- Fluency (or Intelligibility)
- Post-editing effort (time, key strokes, ...)
- ...

Manual vs. 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
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 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 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 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 system improvements

3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems
Problems of Lexical Similarity Measures

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

- Culy and Riehemann [CR03]
- Coughlin [Cou03]
- Callison-Burch et al. [CBOK06]

**Underlying Cause**

Lexical similarity is nor a *sufficient* neither a *necessary* condition so that two sentences convey the same meaning.
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]

→ 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 [CBOK06, KM06]

- 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. Combined Linguistically-motivated Measures

3. Confidence Estimation

4. Conclusions
Can we do better?

Extending Lexical Similarity Measures to increase robustness (avoid sparsity):

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

- Paraphrasing support:
  - Zhou et al. [ZLH06], Kauchak and Barzilay [KB06], Owczarzak et al. [OGGW06]
  - New versions of METEOR, TER
Can we do better?

Extending Lexical Similarity Measures to increase robustness (avoid sparsity):

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

- Paraphrasing support:
  - Zhou et al. [ZLH06], Kauchak and Barzilay [KB06], Owczarzak et al. [OGGW06]
  - New versions of METEOR, TER
Similarity Measures Based on Linguistic Features

More linguistically-motivated measures:

- Features capturing **syntactic** and **semantic** information
- Shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, discourse representation
- Extense bibliography in the last years: [PN07], [LG05], [AGGM06], [MB07] [OvGW07a, OvGW07b], [KSO09], [CN08], [RMDW01], [GM07, GM09], [GMGM10], [PCGJM09], etc.
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
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
Our Approach

(Giménez & Màrquez, 2010)

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

(Almén & Iglesias, 2010)

Combined Linguistically-motivated Measures
Our Approach

(Giménez & Márquez, 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...
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*

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

\(T\): 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
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
Overlap/Matching among Linguistic Elements

- 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)

- Open source software: ASIYA, Open Toolkit for Automatic MT (Meta-)Evaluation (formerly IQ$_{MT}$)
  http://www.lsi.upc.es/~nlp/Asiya/
Overlap/Matching among Linguistic Elements

- 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)

- Open source software: ASIYA, Open Toolkit for Automatic MT (Meta-)Evaluation (formerly IQMT)
  http://www.lsi.upc.es/~nlp/Asiya/
Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]
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>Lexical</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>Syntactic</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>Semantic</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>$\rho_{all}$</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>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><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>Lexical</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>Syntactic</td>
<td>Parts-of-speech</td>
<td></td>
<td></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>Semantic</td>
<td>Semantic Roles</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discourse Repr.</td>
<td>0.72</td>
<td><strong>0.96</strong></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 Similarity
Lexical Recall
Lexical Precision
F-measure
Edit Distance

Syntactic Similarity
PoS Tagging
Dependency Parsing
Chunking
Constituency Parsing
Lemmatization

Semantic Similarity
Named Entities
Semantic Roles
Discourse Representations
Towards Heterogeneous Automatic MT Evaluation
Recent Works on Metric Combination

Different metrics capture different aspects of similarity

Suitable for combination

- Corston-Oliver et al. [COGB01]
- Kulesza and Shieber [KS04]
- Gamon et al. [GAS05]
- Akiba et al. [AIS01]
- Quirk [Qui04]
- Liu and Gildea [LG07]
- Albrecht and Hwa [AH07a]
- Paul et al. [PFS07]
- Ye et al. [YZL07]
- Giménez and Màrquez [GM08]
Recent Works on Metric Combination

Different metrics capture different aspects of similarity

Suitable for combination

- Corston-Oliver et al. [COGB01]
- Kulesza and Shieber [KS04]
- Gamon et al. [GAS05]
- Akiba et al. [AIS01]
- Quirk [Qui04]
- Liu and Gildea [LG07]
- Albrecht and Hwa [AH07a]
- Paul et al. [PFS07]
- Ye et al. [YZL07]
- Giménez and Màrquez [GM08]
The Most Simple Approach: ULC

- Uniformly averaged linear combination of measures (ULC):

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

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

The Most Simple Approach: ULC

- Uniformly averaged linear combination of measures (ULC):

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

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

\[
M = \{ \text{`ROUGE}_W$, `METEOR', `DP-HWC}_r$, `DP-O_c(\star')$, `DP-O_l(\star')$, `DP-O_r(\star')$, `CP-STM_4$, `SR-O_r(\star')$, `SR-O_{rv}$, `DR-O_{rp}(\star') \}
\]
## Evaluation of ULC

WMT 2008 meta-evaluation results (into-English)

<table>
<thead>
<tr>
<th>Measure</th>
<th>$\rho_{sys}$</th>
<th>$\text{consistency}_{snt}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ULC</td>
<td>0.83</td>
<td>0.56</td>
</tr>
<tr>
<td>DP-O$_r$(⋆)</td>
<td>0.83</td>
<td>0.51</td>
</tr>
<tr>
<td>DR-O$_r$(⋆)</td>
<td>0.80</td>
<td>0.50</td>
</tr>
<tr>
<td>METEOR$_{\text{ranking}}$</td>
<td>0.78</td>
<td>0.51</td>
</tr>
<tr>
<td>SR-O$_r$(⋆)</td>
<td>0.77</td>
<td>0.50</td>
</tr>
<tr>
<td>METEOR$_{\text{baseline}}$</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$_{\text{stem+wnsyn}}$</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>$\text{consistency}_{snt}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ULC</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>AE$_{2004}$</th>
<th>CE$_{2004}$</th>
<th>AE$_{2005}$</th>
<th>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$_{ass.}$</td>
<td>5/5</td>
<td>10/10</td>
<td>6/7</td>
<td>5/10</td>
</tr>
<tr>
<td>#sentences$_{ass.}$</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>AE&lt;sub&gt;04&lt;/sub&gt;</th>
<th>CE&lt;sub&gt;04&lt;/sub&gt;</th>
<th>AE&lt;sub&gt;05&lt;/sub&gt;</th>
<th>CE&lt;sub&gt;05&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ULC</strong>&lt;sub&gt;(AE&lt;sub&gt;04&lt;/sub&gt;)&lt;/sub&gt;</td>
<td>0.6392</td>
<td>0.6294</td>
<td>0.5327</td>
<td>0.5695</td>
</tr>
<tr>
<td><strong>ULC</strong>&lt;sub&gt;(CE&lt;sub&gt;04&lt;/sub&gt;)&lt;/sub&gt;</td>
<td>0.6306</td>
<td>0.6333</td>
<td>0.5115</td>
<td>0.5692</td>
</tr>
<tr>
<td><strong>ULC</strong>&lt;sub&gt;(AE&lt;sub&gt;05&lt;/sub&gt;)&lt;/sub&gt;</td>
<td>0.6175</td>
<td>0.6029</td>
<td><strong>0.5450</strong></td>
<td>0.5706</td>
</tr>
<tr>
<td><strong>ULC</strong>&lt;sub&gt;(CE&lt;sub&gt;05&lt;/sub&gt;)&lt;/sub&gt;</td>
<td>0.6218</td>
<td>0.6208</td>
<td>0.5270</td>
<td><strong>0.6047</strong></td>
</tr>
<tr>
<td><strong>Max Indiv.</strong></td>
<td>0.5877</td>
<td>0.5955</td>
<td>0.4960</td>
<td>0.5348</td>
</tr>
</tbody>
</table>
Linguistic Measures at International Campaigns

- **NIST 2004/2005**
  - Arabic-to-English / Chinese-to-English
  - Broadcast news / weblogs / dialogues

- **WMT 2007-2010**
  - Translation between several European languages

- **IWSLT 2005-2008**
  - Spoken language translation
  - Chinese-to-English
Linguistic Measures at International Campaigns

- **NIST 2004/2005**
  - Arabic-to-English / Chinese-to-English
  - Broadcast news / weblogs / dialogues

- **WMT 2007-2010**
  - Translation between several European languages

- **IWSLT 2005-2008**
  - Spoken language translation
  - Chinese-to-English

Controversial results at NIST Metrics MATR08/09 Challenges!
Ongoing and Future Work

1. Metaevaluation of measures
   → Better understand differences between lexical and higher level measures

2. Work on the combination of measures
   → Learning combined similarity measures

3. Porting measures to languages other than English
   → Need of linguistic analyzers

4. Use measures for semi–automatic error analysis
   → (Web) Graphical interface
Ongoing and Future Work

1. Metaevaluation of measures
   → Better understand differences between lexical and higher level measures

2. Work on the combination of measures
   → Learning combined similarity measures

3. Porting measures to languages other than English
   → Need of linguistic analyzers

4. Use measures for semi-automatic error analysis
   → (Web) Graphical interface
Ongoing and Future Work

1. Metaevaluation of measures
   → Better understand differences between lexical and higher level measures

2. Work on the combination of measures
   → Learning combined similarity measures

3. Porting measures to languages other than English
   → Need of linguistic analyzers

4. Use measures for semi-automatic error analysis
   → (Web) Graphical interface
Ongoing and Future Work

1. Metaevaluation of measures
   → Better understand differences between lexical and higher level measures

2. Work on the combination of measures
   → Learning combined similarity measures

3. Porting measures to languages other than English
   → Need of linguistic analyzers

4. Use measures for semi–automatic error analysis
   → (Web) Graphical interface
Talk Overview

1. Automatic MT Evaluation
2. Combined Linguistically-motivated Measures
3. Confidence Estimation
4. Conclusions
Confidence Estimation

New setting:
→ Quality evaluation without reference translations

Motivation:
→ Ranking of several candidate translations when translating new examples

Information available:
→ Source sentence, candidate translation(s), and (possibly) system information
New setting:
→ Quality evaluation without reference translations

Motivation:
→ Ranking of several candidate translations when translating new examples

Information available:
→ Source sentence, candidate translation(s), and (possibly) system information

Johns Hopkins University Summer Workshop, 2003
“Confidence Estimation for Machine Translation” [BFF⁺03]
Confidence Estimation

→ Classification according to the target function

- *Human likeness*
  → discern between human and automatic translations
    - Classification

- *Human acceptability*
  → emulate the behavior of human assessors
    - Classification [GAS05]
    - Linear Regression [Qui04, AH07b, SG10]
    - Ranking [SE10]
Features to train the quality measures:

- System-dependent
- System-independent
Confidence Estimation

**Features** to train the quality 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**
Features to train the quality measures:

- **System-dependent**
- **System-independent**

  → **source** (translation difficulty)
  - sentence length
  - ambiguity → dictionary/alignment/WordNet-based
    (number of candidate translations per word or phrase)

  → **target** (fluency)
  - sentence length
  - language modeling

  → **source-target** (adequacy)
  - length ratio
  - punctuation issues
  - candidate matching → dictionary-/alignment-based
**Features** to train the quality measures:

- System-dependent
- System-independent

**Remark: most valuable features**

- System-dependent
- Based on $n$-best lists
- Capturing target text properties
The FAUST Project (2010-2013)

- Feedback Analysis for User Adaptive Statistical Translation
- Theme FP7-ICT-2009-4
- Objective 2.2: Language-based interaction
- Coordinator: University of Cambridge (Bill Byrne)
- http://divf.eng.cam.ac.uk/faust

**Goal** Develop interactive machine translation systems which adapt rapidly and intelligently to user feedback
CE-related challenge

Create novel automatic metrics of translation quality which reflect preferences learned from user feedback

- State of the art: MT relies on metrics which do not reflect user interest
- FAUST: MT metrics as models of user feedback

Keywords: on-line, adaptive
FAUST: On-line Confidence Estimation
FAUST: On-line Confidence Estimation

**source**

**Ta**
Eric is high

**Tb**
Eric is tall

- [ ] Ta is better than Tb
- [x] Tb is better than Ta
- [ ] Ta and Tb are equally good (or bad)

\[ \text{quality}(Tb) > \text{quality}(Ta) ? \]
FAUST: On-line Confidence Estimation

Ongoing work:

- Preliminary set of 14 CE measures (= features)
- Learn to rank pairwise comparisons
- Ranking perceptron (with linear and polynomial kernels)
- Promising results on an initial batch setting
Talk Overview

1. Automatic MT Evaluation
2. Combined Linguistically-motivated Measures
3. Confidence Estimation
4. Conclusions
Metricwise System Development

Error Detection → Error Analysis → Refinement → Implementation → Test → OK?

MT System Developer

Evaluation Methods

Keep

YES

NO

Discard

Unfruitful Results
Metricwise System Development

Conclusions
Metricwise System Development
Metricwise System Development

- MT System Developer
- Implementation
  - Error Detection
  - Error Analysis
  - Refinement
  - Test
    - OK?
      - YES
      - NO
        - Discard
        - Unfruitful Results

Meta-evaluation
Metricwise System Development

- MT System Developer
- Error Detection
- Error Analysis
- Refinement
- Implementation
- Test
- Meta-evaluation
- Evaluation Methods
- Discard
- Unfruitful Results
- Keep
- YES
- NO
Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. 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)
Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. 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)
Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. 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)
Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. 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)
Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. 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)
Empirical MT is a very active research field

Evaluation methods play a crucial role

Measuring overall translation quality is hard
  → Quality aspects are heterogeneous and diverse

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)
Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. 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)
Automatic Evaluation in Machine Translation
Towards Combined Linguistically-motivated Measures

Lluís Màrquez and Jesús Giménez
TALP Research Center
Technical University of Catalonia

Machine Translation and Morphologically-rich Languages
Research Workshop of the Israel Science Foundation
University of Haifa, January 24, 2010
### On-line Confidence Estimation

Preliminary set of features

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE-BiDictO</td>
<td>bilingual dictionary based overlap</td>
</tr>
<tr>
<td>CE-(N_c)</td>
<td>source/candidate phrase chunk ratio</td>
</tr>
<tr>
<td>CE-(N_e)</td>
<td>source/candidate named entity ratio</td>
</tr>
<tr>
<td>CE-(O_c)</td>
<td>source/candidate phrase chunk overlap</td>
</tr>
<tr>
<td>CE-(O_e)</td>
<td>source/candidate named entity overlap</td>
</tr>
<tr>
<td>CE-(O_p)</td>
<td>source/candidate part-of-speech overlap</td>
</tr>
<tr>
<td>CE-ippl</td>
<td>candidate language model inverse perplexity</td>
</tr>
<tr>
<td>CE-ippl(_C)</td>
<td>candidate chunk language model inverse perplexity</td>
</tr>
<tr>
<td>CE-ippl(_P)</td>
<td>candidate PoS language model inverse perplexity</td>
</tr>
<tr>
<td>CE-length</td>
<td>source/candidate length ratio</td>
</tr>
<tr>
<td>CE-long</td>
<td>source/candidate length ratio (penalize short candidates)</td>
</tr>
<tr>
<td>CE-oov</td>
<td>candidate language model out-of-vocabulary tokens ratio</td>
</tr>
<tr>
<td>CE-short</td>
<td>source/candidate length ratio (penalize long candidates)</td>
</tr>
<tr>
<td>CE-symbols</td>
<td>symbol overlap (punctuation, etc.)</td>
</tr>
</tbody>
</table>
Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez.
MT Evaluation: Human-Like vs. Human Acceptable.

Joshua Albrecht and Rebecca Hwa.
A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation.

Joshua Albrecht and Rebecca Hwa.
Regression for Sentence-Level MT Evaluation with Pseudo References.


Conclusions

Jesús Giménez and Lluís Màrquez.
Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems.

Jesús Giménez and Lluís Màrquez.
Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations.

Jesús Giménez and Lluís Màrquez.
On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation.

David Kauchak and Regina Barzilay.
Paraphrasing for Automatic Evaluation.


Ding Liu and Daniel Gildea.  

Dennis Mehay and Chris Brew.  
BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation.  

Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way.
Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation.


Karolina Owczarzak, Josef van Genabith, and Andy Way.

Dependency-Based Automatic Evaluation for Machine Translation.


Karolina Owczarzak, Josef van Genabith, and Andy Way.

Labelled Dependencies in Machine Translation Evaluation.


Michael Paul, Andrew Finch, and Eiichiro Sumita.
Reducing Human Assessments of Machine Translation Quality to Binary Classifiers.


**Maja Popovic and Hermann Ney.**

Word Error Rates: Decomposition over POS classes and Applications for Error Analysis.


**Chris Quirk.**

Training a Sentence-Level Machine Translation Confidence Metric.

Florence Reeder, Keith Miller, Jennifer Doyon, and John White.
The Naming of Things and the Confusion of Tongues: an MT Metric.

Radu Soricut and Abdessamad Echihabi.
Trustrank: Inducing trust in automatic translations via ranking.

Lucia Specia and Jesús Giménez.
Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation.

Yang Ye, Ming Zhou, and Chin-Yew Lin.
Sentence Level Machine Translation Evaluation as a Ranking.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy.
Re-evaluating Machine Translation Results with Paraphrase Support.