Towards Heterogeneous Automatic MT Error Analysis

(6th LREC)

Jesús Giménez and Lluís Màrquez

TALP Research Center
Technical University of Catalonia

May 29, 2008
1. Introduction
2. Our Proposal
3. Applicability
4. Discussion
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Introduction

Outline

1 Introduction
   - The Role of Evaluation Methods
   - Recent Advances in Automatic MT Evaluation

2 Our Proposal

3 Applicability

4 Discussion
Towards Heterogeneous Automatic MT Error Analysis (*6th LREC*)

Outline

1. Introduction
   - The Role of Evaluation Methods
   - Recent Advances in Automatic MT Evaluation

2. Our Proposal

3. Applicability

4. Discussion
Development Cycle of MT systems

1. MT System Developer:
   - Identify a Type of Error
   - Analyze Possible Causes
   - Suggest a Mechanism for Improvement
   - Implement it
   - Test it
     - YES: then keep it
     - NO: then discard it
     - unfruitful results

   - then keep it
   - then discard it
Towards Heterogeneous Automatic MT Error Analysis (*6th LREC*)

Introduction

The Role of Evaluation Methods

Development Cycle of MT systems
Error Analysis Today

- Error analyses are conducted manually
  - low-level analysis related to the linguistic analysis of translation quality (i.e., what?)
  - high-level analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations
Introduction

The Role of Evaluation Methods

Error Analysis Today

- Error analyses are conducted manually
  - low-level analysis related to the linguistic analysis of translation quality (i.e., what?)
  - high-level analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations
Error Analysis Today

- Error analyses are conducted manually
  - low-level analysis related to the linguistic analysis of translation quality (i.e., what?)
  - high-level analysis involving knowledge about the system architecture (i.e., why?)

- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations
Error Analysis Today

- Error analyses are conducted manually
  - **low-level** analysis related to the linguistic analysis of translation quality (i.e., what?)
  - **high-level** analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
  - Automatic metrics are used only as quantitative evaluation measures
    - to identify high/low quality translations
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Error Analysis Today

- Error analyses are conducted manually
  - low-level analysis related to the linguistic analysis of translation quality (i.e., what?)
  - high-level analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Introduction

The Role of Evaluation Methods

Metrics Based on Lexical Similarity

- **Edit Distance**
  - WER, PER, TER

- **Precision**
  - BLEU, NIST, WNM

- **Recall**
  - ROUGE, CDER

- **Precision/Recall**
  - GTM, METEOR, BLANC, SIA
Outline

1. Introduction
   - The Role of Evaluation Methods
   - Recent Advances in Automatic MT Evaluation

2. Our Proposal

3. Applicability

4. Discussion
Extending the Reference Lexicon

- Lexical variants
  - Morphological variations (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]
Beyond the Lexical Level

Syntactic Similarity

- Shallow Parsing
  - Popovic and Ney [PN07]
  - Giménez and Màrquez [GM07]

- Constituency Parsing
  - Liu and Gildea [LG05]

- Dependency Parsing
  - Liu and Gildea [LG05]
  - Amigó et al. [AGGM06]
  - Mehay and Brew [MB07]
  - Owczarzak et al. [OvGW07]
Beyond the Lexical Level

Semantic Similarity

- Semantic Roles
  - Giménez and Màrquez [GM07]

- Named Entities
  - Reeder et al. [RMDW01]
  - Giménez and Màrquez [GM07]

- Discourse Representations
  - Giménez and Màrquez [GM08b]
Outline

1. Introduction

2. Our Proposal
   - A Smorgasbord of Features

3. Applicability

4. Discussion
Our Proposal

Rely on Automatic Metrics

**Idea:** Let automatic metrics do most of the *low-level* analysis, so system developers may concentrate on *high-level* analysis.
Towards Heterogeneous Automatic MT Error Analysis *(6th LREC)*

Our Proposal

Heterogeneous Error Analysis

- as automatic as possible
- as heterogeneous as possible
  - Quality Aspects: lexical, syntactic, semantic, etc.
  - Granularity
    - fine aspects → transfer of specific linguistic elements (e.g., what proportion of singular nouns are correctly translated?)
    - coarse aspects → overall linguistic structure (e.g., what proportion of the semantic role structure is correctly translated?)
Heterogeneous Error Analysis

- as automatic as possible
- as heterogeneous as possible
  - Quality Aspects: lexical, syntactic, semantic, etc.
  - Granularity
    - fine aspects → transfer of specific linguistic elements
      (e.g., what proportion of singular nouns are correctly translated?)
    - coarse aspects → overall linguistic structure
      (e.g., what proportion of the semantic role structure is correctly translated?)
Towards Heterogeneous Automatic MT Error Analysis (*6th LREC*)

---

**Our Proposal**

**Heterogeneous Error Analysis**

- as automatic as possible
- as heterogeneous as possible
  - **Quality Aspects**: lexical, syntactic, semantic, etc.
  - **Granularity**
    - fine aspects → transfer of specific linguistic elements (e.g., what proportion of singular nouns are correctly translated?)
    - coarse aspects → overall linguistic structure (e.g., what proportion of the semantic role structure is correctly translated?)
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Our Proposal

Heterogeneous Error Analysis

- as automatic as possible
- as heterogeneous as possible
  - **Quality Aspects:** lexical, syntactic, semantic, etc.
  - **Granularity**
    - fine aspects → transfer of specific linguistic elements
      (e.g., what proportion of singular nouns are correctly translated?)
    - coarse aspects → overall linguistic structure
      (e.g., what proportion of the semantic role structure is correctly translated?)
Towards Heterogeneous Automatic MT Error Analysis (*6th LREC*)

Our Proposal

A Smorgasbord of Features

Outline

1. Introduction
2. Our Proposal
   - A Smorgasbord of Features
3. Applicability
4. Discussion
Linguistic Similarities

More than 500 metric variants operating at different linguistic levels:

- Lexical
- Shallow Syntactic (Lemmatization, PoS Tagging, and Base Phrase Chunking)
- Syntactic (Constituency and Dependency Parsing)
- Shallow Semantic (Semantic Roles and Named Entities)
- Semantic (Discourse Representations)
Shallow Syntactic Level

**SP-O_p-★** Average overlapping between words belonging to the same PoS.

**SP-O_c-★** Average overlapping between words belonging to the same phrase chunk type.

**SP-NIST_l** NIST score over sequences of lemmas.

**SP-NIST_p** NIST score over PoS sequences.

**SP-NIST_{iob}** NIST score over chunk IOB sequences.

**SP-NIST_c** NIST score over sequences of chunks.
Our Proposal
A Smorgasbord of Features

Syntactic Level (i)

- **Dependency Overlapping**
  - **DP-O₁-⋆** Average overlapping between words hanging at the same level.
  - **DP-Oᶜ-⋆** Average overlapping between words hanging from terminal nodes (i.e., grammatical categories).
  - **DP-Oᵣ-⋆** Average overlapping between words ruled by non-terminal nodes (i.e., grammatical relations).
Syntactic Level (ii)

- **Head-word Chain Matching** (Liu and Gildea [LG05])
  - $\text{DP-HWC}_w$ Average head-word chain matching up to length-4 word chains.
  - $\text{DP-HWC}_c$ Average head-word chain matching up to length-4 category chains.
  - $\text{DP-HWC}_r$ Average head-word chain matching up to length-4 relation chains.
Syntactic Level (iii)

- **Syntactic Overlapping**
  - **CP-O$_p$-$\star$** Average overlapping between words belonging to the same PoS (similar to ‘SP-O$_p$-$\star$’).
  - **CP-O$_c$-$\star$** Average overlapping between words belonging to the same phrase type (similar to ‘SP-O$_c$-$\star$’).

- **Syntactic Tree Matching** (Liu and Gildea [LG05])
  - **CP-STM** Constituent tree matching averaged up to length-9 syntactic subpaths.
Shallow Semantic Level (i)

- **Named Entity Overlapping/Matching**
  - **NE-O\textsubscript{e-\star}**: Average lexical overlapping between named entities of the same type (excluding type ‘O’, i.e., Not-a-NE).
  - **NE-O\textsubscript{e-\star\star}**: Average lexical overlapping between named entities of the same type (including ‘O’).
  - **NE-M\textsubscript{e-\star}**: Average lexical matching between named entities of the same type.
Semantic Role Overlapping/Matching

**SR-O_r-✩** Average lexical overlapping between semantic roles (arguments and adjuncts) of the same type.

**SR-M_r-✩** Average lexical matching between semantic roles of the same type.

**SR-O_r** Role overlapping independently from the lexical realization.
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Our Proposal

A Smorgasbord of Features

Semantic Level

- **Discourse Overlapping**
  - **DR-O_{or}** Average lexical overlapping between DR structures of the same type.
  - **DR-O_{rp}** Average morphosyntactic overlapping between DR structures of the same type.

- **Semantic Tree Matching**
  - **DR-STM** Matching between discourse representations averaged up to length-9 semantic subpaths.
Linguistic Features at Work

ACL’07 MT Workshop (French/German/Spanish/Czech-to-English)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Adeq.</th>
<th>Fluen.</th>
<th>Rank</th>
<th>Const.</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR-O&lt;sub&gt;r&lt;/sub&gt;-⋆</td>
<td>.774</td>
<td>.839</td>
<td>.803</td>
<td>.741</td>
<td>.789</td>
</tr>
<tr>
<td>ParaEval-Recall</td>
<td>.712</td>
<td>.742</td>
<td>.768</td>
<td>.798</td>
<td>.755</td>
</tr>
<tr>
<td>METEOR</td>
<td>.701</td>
<td>.719</td>
<td>.745</td>
<td>.669</td>
<td>.709</td>
</tr>
<tr>
<td>BLEU</td>
<td>.690</td>
<td>.722</td>
<td>.672</td>
<td>.602</td>
<td>.671</td>
</tr>
<tr>
<td>1-TER</td>
<td>.607</td>
<td>.538</td>
<td>.520</td>
<td>.514</td>
<td>.644</td>
</tr>
<tr>
<td>Max Fluen. Corr.</td>
<td>.644</td>
<td>.653</td>
<td>.656</td>
<td>.512</td>
<td>.616</td>
</tr>
<tr>
<td>GTM</td>
<td>.655</td>
<td>.674</td>
<td>.616</td>
<td>.495</td>
<td>.610</td>
</tr>
</tbody>
</table>
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Applicability

Outline

1. Introduction

2. Our Proposal

3. Applicability
   - Settings
   - Document Level Error Analysis
   - Sentence Level Error Analysis

4. Discussion
Outline

1. Introduction
2. Our Proposal
3. Applicability
   - Settings
     - Document Level Error Analysis
     - Sentence Level Error Analysis
4. Discussion
NIST 2005 MT Evaluation Puzzle

Arabic-to-English Translation Exercise [LP05]
Linguistic Features Solved the Puzzle

- Giménez and Màrquez [GM07]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>$R_{sys}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td>BLEU</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>GTM</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Syntactic</strong></td>
<td>SP-NIST$_p$</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>DP-HWC$_r$</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>CP-STM</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>Semantic</strong></td>
<td>SR-$O_{r-\star}$</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>SR-$M_{r-\star}$</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>DR-$O_{r-\star}$</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>DR-$O_{rp-\star}$</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Linguistic Features Solved the Puzzle

- Giménez and Màrquez [GM07]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>$R_{sys}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td>BLEU</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>GTM</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Syntactic</strong></td>
<td>SP-NIST$_p$</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>DP-HWC$_r$</td>
<td><strong>0.88</strong></td>
</tr>
<tr>
<td></td>
<td>CP-STM</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>Semantic</strong></td>
<td>SR-$O_r$-$\star$</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>SR-$M_r$-$\star$</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>DR-$O_r$-$\star$</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>DR-$O_{rp}$-$\star$</td>
<td><strong>0.97</strong></td>
</tr>
</tbody>
</table>
A Note on Meta-Evaluation

- Metrics are automatically evaluated on the basis of **human likeness**, i.e., in terms of their ability to distinguish manual from automatic translations.
  - ORANGE, Lin and Och [LO04]
  - KING, Amigó et al. [AGPV05]

- We use the KING measure
  - “A metric should never rank any reference translation lower in quality than any automatic translation.”

- KING(\(x\)) serves as an estimate of the impact on system performance of the quality aspects captured by metric \(x\)
### Lexical Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>KING</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit Distance</td>
<td>1-PER</td>
<td>0.63</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>1-TER</td>
<td>0.70</td>
<td>0.53</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>1-WER</td>
<td>0.67</td>
<td>0.49</td>
<td>0.54</td>
</tr>
<tr>
<td>Precision</td>
<td>BLEU</td>
<td>0.65</td>
<td>0.47</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>NIST</td>
<td>0.69</td>
<td>10.63</td>
<td>11.27</td>
</tr>
<tr>
<td>Recall</td>
<td>ROUGE$_W$</td>
<td>0.68</td>
<td>0.31</td>
<td>0.33</td>
</tr>
<tr>
<td>F-measure</td>
<td>GTM ($e = 1$)</td>
<td>0.64</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>GTM ($e = 2$)</td>
<td>0.66</td>
<td>0.31</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>METEOR$_{exact}$</td>
<td>0.68</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>METEOR$_{wnsyn}$</td>
<td>0.68</td>
<td>0.64</td>
<td>0.68</td>
</tr>
</tbody>
</table>
## Shallow Syntactic Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>KING</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PoS</strong></td>
<td>SP-(O_p)-(\star)</td>
<td>0.64</td>
<td>0.52</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>SP-(O_p)-J</td>
<td>0.26</td>
<td>0.53</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>SP-(O_p)-N</td>
<td>0.53</td>
<td>0.57</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>SP-(O_p)-V</td>
<td>0.43</td>
<td>0.39</td>
<td>0.41</td>
</tr>
<tr>
<td><strong>Overlapping</strong></td>
<td>SP-(O_c)-(\star)</td>
<td>0.63</td>
<td>0.54</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>Chunk</strong></td>
<td>SP-(O_c)-NP</td>
<td>0.60</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Overlapping</strong></td>
<td>SP-(O_c)-PP</td>
<td>0.38</td>
<td>0.63</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>SP-(O_c)-VP</td>
<td>0.41</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td><strong>NIST(_x)</strong></td>
<td>SP-NIST(_l)-5</td>
<td>0.69</td>
<td>10.78</td>
<td>11.44</td>
</tr>
<tr>
<td></td>
<td>SP-NIST(_p)-5</td>
<td>0.71</td>
<td>8.74</td>
<td>9.04</td>
</tr>
<tr>
<td></td>
<td>SP-NIST(_iob)-5</td>
<td>0.65</td>
<td>6.81</td>
<td>6.91</td>
</tr>
<tr>
<td></td>
<td>SP-NIST(_c)-5</td>
<td>0.57</td>
<td>6.13</td>
<td>6.18</td>
</tr>
</tbody>
</table>
### Syntactic Features (i)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>KING</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP-HWC&lt;sub&gt;W&lt;/sub&gt;-4</td>
<td>0.59</td>
<td>0.14</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>DP-HWC&lt;sub&gt;C&lt;/sub&gt;-4</td>
<td>0.48</td>
<td>0.42</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>DP-HWC&lt;sub&gt;R&lt;/sub&gt;-4</td>
<td>0.52</td>
<td>0.33</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>DP-O&lt;sub&gt;I&lt;/sub&gt;-a</td>
<td>0.58</td>
<td>0.41</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>DP-O&lt;sub&gt;C&lt;/sub&gt;-a</td>
<td>0.60</td>
<td>0.50</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>DP-O&lt;sub&gt;C&lt;/sub&gt;-aux</td>
<td>0.30</td>
<td>0.51</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>DP-O&lt;sub&gt;C&lt;/sub&gt;-det</td>
<td>0.14</td>
<td>0.56</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>DP-O&lt;sub&gt;C&lt;/sub&gt;-n</td>
<td>0.35</td>
<td>0.75</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>DP-O&lt;sub&gt;C&lt;/sub&gt;-v</td>
<td>0.57</td>
<td>0.57</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>Dependency Parsing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DP-O&lt;sub&gt;C&lt;/sub&gt;-a</td>
<td>0.37</td>
<td>0.43</td>
<td>0.45</td>
<td></td>
</tr>
</tbody>
</table>
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Applicability

Document Level Error Analysis

Syntactic Features (ii)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>KING</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DP-(O_r)-*</td>
<td></td>
<td>0.66</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>DP-(O_r)-aux</td>
<td></td>
<td>0.14</td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>DP-(O_r)-det</td>
<td></td>
<td>0.35</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>DP-(O_r)-fc</td>
<td></td>
<td>0.21</td>
<td>0.26</td>
<td>0.24</td>
</tr>
<tr>
<td>DP-(O_r)-i</td>
<td></td>
<td>0.60</td>
<td>0.44</td>
<td>0.43</td>
</tr>
<tr>
<td>DP-(O_r)-obj</td>
<td></td>
<td>0.43</td>
<td>0.36</td>
<td>0.35</td>
</tr>
<tr>
<td>DP-(O_r)-s</td>
<td></td>
<td>0.47</td>
<td>0.52</td>
<td>0.45</td>
</tr>
<tr>
<td><strong>Constituency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP-(O_p)-*</td>
<td></td>
<td>0.64</td>
<td>0.52</td>
<td>0.55</td>
</tr>
<tr>
<td>CP-(O_c)-*</td>
<td></td>
<td>0.63</td>
<td>0.50</td>
<td>0.53</td>
</tr>
<tr>
<td>CP-(O_c)-NP</td>
<td></td>
<td>0.61</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>CP-(O_c)-PP</td>
<td></td>
<td>0.51</td>
<td>0.50</td>
<td>0.53</td>
</tr>
<tr>
<td>CP-(O_c)-SBAR</td>
<td></td>
<td>0.36</td>
<td>0.36</td>
<td>0.38</td>
</tr>
<tr>
<td>CP-STM-9</td>
<td></td>
<td>0.58</td>
<td>0.35</td>
<td>0.35</td>
</tr>
</tbody>
</table>
## Shallow Semantic Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>KING</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Named Entities</strong></td>
<td>NE-M$_e$-⋆</td>
<td>0.32</td>
<td>0.53</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>NE-M$_e$-ORG</td>
<td>0.11</td>
<td>0.27</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>NE-M$_e$-PER</td>
<td>0.13</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td><strong>Semantic Roles</strong></td>
<td>SR-M$_r$-⋆</td>
<td>0.50</td>
<td><strong>0.19</strong></td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>SR-M$_r$-A0</td>
<td>0.33</td>
<td><strong>0.31</strong></td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>SR-M$_r$-A1</td>
<td>0.28</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>SR-O$_r$</td>
<td>0.41</td>
<td><strong>0.64</strong></td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>SR-O$_r$-⋆</td>
<td><strong>0.53</strong></td>
<td>0.36</td>
<td><strong>0.37</strong></td>
</tr>
<tr>
<td></td>
<td>SR-O$_r$-AM-TMP</td>
<td>0.13</td>
<td><strong>0.39</strong></td>
<td>0.38</td>
</tr>
</tbody>
</table>
## Semantic Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>KING</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse Representations</td>
<td>DR-O&lt;sub&gt;r&lt;/sub&gt;-**</td>
<td>0.59</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>DR-O&lt;sub&gt;r&lt;/sub&gt;-card</td>
<td>0.12</td>
<td>0.49</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>DR-O&lt;sub&gt;r&lt;/sub&gt;-dr</td>
<td>0.56</td>
<td>0.43</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>DR-O&lt;sub&gt;r&lt;/sub&gt;-eq</td>
<td>0.12</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>DR-O&lt;sub&gt;r&lt;/sub&gt;-named</td>
<td>0.38</td>
<td>0.48</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>DR-O&lt;sub&gt;r&lt;/sub&gt;-pred</td>
<td>0.55</td>
<td>0.38</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>DR-O&lt;sub&gt;r&lt;/sub&gt;-prop</td>
<td>0.39</td>
<td>0.27</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>DR-O&lt;sub&gt;r&lt;/sub&gt;-rel</td>
<td>0.56</td>
<td>0.38</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>DR-STM-9</td>
<td>0.40</td>
<td>0.26</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. Our Proposal
3. Applicability
   - Settings
   - Document Level Error Analysis
   - Sentence Level Error Analysis
4. Discussion
### Ex: Thousand Monks

<table>
<thead>
<tr>
<th>Ref</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>Over 1000 monks and nuns, observers and scientists from over 30 countries and the host country attended the religious summit held for the first time in Myanmar which started today, Thursday.</td>
</tr>
<tr>
<td>2:</td>
<td>More than 1000 monks, nuns, observers and scholars from more than 30 countries, including the host country, participated in the religious summit which Myanmar hosted for the first time and which began on Thursday.</td>
</tr>
<tr>
<td>3:</td>
<td>The religious summit, staged by Myanmar for the first time and began on Thursday, was attended by over 1,000 monks an nuns, observers and scholars from more than 30 countries and host Myanmar.</td>
</tr>
<tr>
<td>4:</td>
<td>More than 1,000 monks, nuns, observers and scholars from more than 30 countries and the host country Myanmar participated in the religious summit, which is hosted by Myanmar for the first time and which began on Thursday.</td>
</tr>
<tr>
<td>5:</td>
<td>The religious summit, which started on Thursday and was hosted for the first time by Myanmar, was attended by over 1,000 monks and nuns, observers and scholars from more than 30 countries and the host country Myanmar.</td>
</tr>
</tbody>
</table>
**Ex: Thousand Monks**

| **Info:** | (1) → **subject:** over/more_than 1,000 monks and nuns, observers and scientists/scholars from over/more_than 30 countries, and/including the host country **action:** attended/participated_in  
(2) → **subject:** the religious summit **action:** began/started  
**temporal:** on Thursday  
(3) → **object:** the religious summit **action:** hosted  
**subject:** by Myanmar **mode:** for the first time |
| **LinearB:** | 1000 monks from more than 30 States and the host State Myanmar attended the Summit, which began on Thursday, hosted by Myanmar for the first time. |
| **Best SMT:** | Religious participated in the summit, hosted by Myanmar for the first time began on Thursday, as an observer and the world of the 1000 monk nun from more than 30 countries and the host state Myanmar. |
### Ex: Thousand Monks - Lexical Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human</strong></td>
<td>Adequacy</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Fluency</td>
<td>3.5</td>
<td>2</td>
</tr>
<tr>
<td><strong>Edit Distance</strong></td>
<td>1-PER</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>1-TER</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>1-WER</td>
<td>0.40</td>
<td>0.48</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>BLEU</td>
<td>0.44</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>NIST</td>
<td>9.04</td>
<td>9.96</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>ROUGE_W</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td><strong>F-measure</strong></td>
<td>GTM (e = 2)</td>
<td>0.30</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>METEOR_wnsyn</td>
<td>0.59</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Applicability

Sentence Level Error Analysis

Ex: Thousand Monks - Shallow Syntactic Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PoS Overlapping</td>
<td>SP-$O_p$-*</td>
<td>0.52</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>SP-$O_p$-IN</td>
<td>0.71</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>SP-$O_p$-NN</td>
<td>0.67</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>SP-$O_p$-NNP</td>
<td>0.60</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>SP-$O_p$-V</td>
<td>0.40</td>
<td>0.75</td>
</tr>
<tr>
<td>Chunk Overlapping</td>
<td>SP-$O_c$-*</td>
<td>0.56</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>SP-$O_c$-NP</td>
<td>0.56</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>SP-$O_c$-PP</td>
<td>0.80</td>
<td>0.71</td>
</tr>
<tr>
<td>NIST$_x$</td>
<td>SP-NIST$_p$</td>
<td>6.21</td>
<td>8.36</td>
</tr>
<tr>
<td></td>
<td>SP-NIST$_c$</td>
<td>6.43</td>
<td>6.25</td>
</tr>
<tr>
<td></td>
<td>SP-NIST$_{iob}$</td>
<td>5.78</td>
<td>6.41</td>
</tr>
</tbody>
</table>
## Applicability

### Sentence Level Error Analysis

**Ex: Thousand Monks - Syntactic Features**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependency</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsing</td>
<td>DP-HWC(w)-4</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>DP-(O_r)-(\star)</td>
<td>0.46</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>DP-(O_r)-mod</td>
<td>0.62</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>DP-(O_r)-obj</td>
<td>0.29</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>DP-(O_r)-pcomp-n</td>
<td>0.71</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>DP-(O_r)-rel</td>
<td>0.33</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Constituency</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsing</td>
<td>CP-(O_c)-(\star)</td>
<td>0.59</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>CP-(O_c)-NP</td>
<td>0.59</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>CP-(O_c)-PP</td>
<td>0.57</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>CP-(O_c)-SB</td>
<td>0.73</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>CP-(O_c)-VP</td>
<td>0.64</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>CP-STM-9</td>
<td>0.34</td>
<td>0.23</td>
</tr>
</tbody>
</table>
### Ex: Thousand Monks - Semantic Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>LinearB</th>
<th>Best SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Semantic Roles</strong></td>
<td>SR-$O_r$</td>
<td>0.84</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>SR-$O_r$-$\ast$</td>
<td>0.56</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>SR-$O_r$-A0</td>
<td>0.44</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>SR-$O_r$-A1</td>
<td>0.57</td>
<td>0.28</td>
</tr>
<tr>
<td><strong>Discourse Representations</strong></td>
<td>DR-$O_r$-$\ast$</td>
<td>0.45</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>DR-$O_r$-dr</td>
<td>0.57</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>DR-$O_r$-nam</td>
<td>0.75</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>DR-$O_r$-pred</td>
<td>0.44</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>DR-$O_r$-rel</td>
<td>0.51</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>DR-STM-9</td>
<td>0.32</td>
<td>0.29</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. Our Proposal
3. Applicability
4. Discussion
   - Conclusions
   - Future Work
Outline

1. Introduction
2. Our Proposal
3. Applicability
4. Discussion
   - Conclusions
   - Future Work
We have presented a valid path towards **heterogeneous automatic MT error analysis**:  
- Our approach allows developers to rapidly obtain detailed qualitative linguistic reports on their system’s capabilities.  
- Human efforts may concentrate on high-level analysis.
Hey! Linguistic Metrics are Not the Panacea

Linguistic metrics rely on:
1. the representativity of the set of human references
   - lexicon
   - grammar
   - style...
2. automatic linguistic processors are
   - domain-dependent
   - language-dependent
   - prone to error
   - slow

Sentence level performance must be improved!

---

1 Panacea: a remedy for all ills or difficulties (see cure-all).
Hey! Linguistic Metrics are Not the Panacea

Linguistic metrics rely on:

1. the representativity of the set of human references
   - lexicon
   - grammar
   - style...

2. automatic linguistic processors are
   - domain-dependent
   - language-dependent
   - prone to error
   - slow

Sentence level performance must be improved!

---

1 Panacea: a remedy for all ills or difficulties (see cure-all).
Outline

1. Introduction
2. Our Proposal
3. Applicability
4. Discussion
   - Conclusions
   - Future Work
Ongoing Steps...

1. Improving sentence level behavior:
   - Backing off to lexical similarity [GM08b]
   - Working on metric combinations [GM08a]

2. Porting metrics to languages other than English (e.g., Castilian Spanish, Catalan...)
A New Interface
Thanks for your Attention

\[ \text{IQ}_{\text{MT}} \text{ v2.0 is publicly available at:} \]

http://www.lsi.upc.edu/~nlp/IQMT
Towards Heterogeneous Automatic MT Error Analysis (6th LREC)

Discussion

Future Work

Towards Heterogeneous Automatic MT Error Analysis

(6th LREC)

Jesús Giménez and Lluís Màrquez

TALP Research Center
Technical University of Catalonia

May 29, 2008
Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez.
MT Evaluation: Human-Like vs. Human Acceptable.
In *Proceedings of COLING-ACL06*, 2006.

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo.

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 Linguistic 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.


Maja Popovic and Hermann Ney.
Discussion

Future Work

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

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

Liang Zhou, Chin-Yew Lin, and Eduard Hovy.
Re-evaluating Machine Translation Results with Paraphrase Support.
Towards Heterogeneous Automatic MT Error Analysis (*6th LREC*)

Discussion

Future Work