A Web Application for the Diagnostic Evaluation of Machine Translation over Specific Linguistic Phenomena

Antonio Toral, Sudip Kumar Naskar, Joris Vreeke, Federico Gaspari, Declan Groves

CNGL & School of Computing, Dublin City University, Ireland

Motivation

- State-of-the-art methods for automatic MT evaluation are typically represented by BLEU, NIST, METEOR, etc.
- These metrics do not provide any feedback or detailed analysis of how an MT system performs on relevant linguistic units
- Diagnostic evaluation provides quantitative analysis as well as qualitative feedback of MT output

Benefits

- A checkpoint is a linguistically-motivated unit (e.g. an ambiguous word, a noun phrase, a prepositional phrase etc.), predefined in a linguistic taxonomy
- Diagnostic evaluation provides qualitative linguistic feedback to MT developers to help them improve the performance of systems
- SOAP-compliant web service built on top of open-source software
- Web-based application provides a very user-friendly evaluation environment; no installation, configuration or maintenance required
- Easily adaptable to new language pairs (in contrast to existing diagnostic tools e.g. Woodpecker (Zhou et al., 2008))
- Low-cost, efficient document-level scoring

DELIC4MT

- Two web-based components: web service (via Panacea platform) and web application
- Occurrences of defined linguistic phenomena are identified in source side of test set
- Using pre-computed word alignments, equivalent tokens of these occurrences are extracted from target side
- For each checkpoint instance, tool checks how many n-grams present in the reference of the checkpoint instance are contained in the MT output

Correlation Results

- DELIC4MT document-level scores correlate with human evaluation scores, in terms of adequacy and fluency, as well as with traditional automatic metrics

<table>
<thead>
<tr>
<th></th>
<th>Eval 1</th>
<th>Eval 2</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>DELIC4MT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(manual alignments)</td>
<td>0.975</td>
<td>0.972</td>
<td>0.972</td>
<td>-0.986</td>
</tr>
<tr>
<td>(auto. alignments)</td>
<td>0.980</td>
<td>0.978</td>
<td>0.966</td>
<td>-0.990</td>
</tr>
</tbody>
</table>

Availability and Future Work

- Freely available for use at: http://tinyurl.com/delic4mt
- Provision of more detailed evaluation reporting
- Integration with existing automatic evaluation metrics (BLEU, TER, METEOR) to create a more comprehensive web-based suite of evaluation tools
- Developed as part of the CoSyne (www.cosyne.eu) and Abu-MaTran (www.abumatran.eu) projects and CNGL (www.cngl.org)

Publications

