A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints

Sudip Kumar Naskar†‡, Antonio Toral†, Federico Gaspari†, and Andy Way*

NCLT† / CNGL‡ School of Computing Dublin City University Dublin 9, Ireland
{snaskar, atoral, fgaspari}@computing.dcu.ie

Applied Language Solutions*
Huddersfield Road
Delph, UK
andy.way@appliedlanguage.com
Outline

- Introduction
- Linguistic checkpoints
- Methodology
- Evaluation
- Conclusions
- Future work
Introduction

Automatic MT evaluation
• Metrics (e.g. BLEU) provide overall score → opaque
• Diagnostic evaluation → identify particular problems
  – Components of MT system
  – Input (e.g. linguistic phenomena)
Introduction: CoSyne

• Consortium
  • Academic / research
  • Users

• Aim: synchronisation of multilingual wikis using MT and other techniques

• Diagnostic Evaluation
  • Feedback to MT developers on system's weaknesses
Linguistic Checkpoints

- Linguistically-motivated units, e.g. ambiguous words, verb-object collocations
- Make up a taxonomy: inventory of linguistic phenomena of the source language that can present problems
  - inherent ambiguity
  - translation into a specific target language, e.g. syntactic divergence between the two languages
- Form the basis of linguistic test suites, the means by which the MT output is evaluated
Linguistic Checkpoints

- Zhou et al. (2008) implemented a tool which:
  - Automatically extracts checkpoints using PoS taggers, word aligners and parsers.
  - Performs n-gram-based evaluation of the checkpoints.
- It has two important limitations
  - Language-dependent data for English–Chinese (the language pair considered in their paper) is hardcoded.
  - Restrictive license (MSR-LA)
Methodology

DELiC4MT (Diagnostic Evaluation using Linguistic Checkpoints For Machine Translation)
http://www.computing.dcu.ie/~atoral/delic4mt/

Makes extensive use of available components and standards:

- Uses state-of-the-art PoS taggers and word aligners
- Exploits the Travelling Object (TO, ~XCES) format to represent word alignment
- Uses the KYOTO Annotation Format (KAF) to represent textual analysis
- Makes use of Kybots to define the evaluation targets (linguistic checkpoints)
Methodology

• Benefits of DELiC4MT

  • End-users can create new evaluation targets for any new language pair

    provided that the phenomena of interest are covered by the linguistic analysis available

  • It can work with any PoS tagger / word aligner

    provided that their output can be converted to the KAF and TO formats, respectively.

  • Takes advantage of the outcomes of recently completed and ongoing FP7 projects
Methodology
Methodology

Diagnostic Evaluation with Linguistic Checkpoints
Methodology
Methodology

Diagnostic Evaluation with Linguistic Checkpoints
Methodology

Diagnostic Evaluation with Linguistic Checkpoints
Text Analysis

Sample KAF files for the Italian-English sentence pair

“È difficile rispondere”

“That is hard to answer”

KEN

<text>
[...]  
<wf wid="w962_1" sent="962" para="1">È</wf>  
<wf wid="w962_2" sent="962" para="1">difficile</wf>  
<wf wid="w962_3" sent="962" para="1">rispondere</wf>  
[...]  
</text>

<terms>
[...]  
<term tid="t962_1" lemma="essere" pos="VER:pres">  
<target id="w962_1"/> </term>  
[...]  
</terms>

<text>
<term tid="t962_2" lemma="difficile" pos="ADJ">  
<target id="w962_2"/> </term>  
<term tid="t962_3" lemma="rispondere" pos="VER:infi">  
<target id="w962_3"/> </term>  
[...]  
</text>

<terms>
[...]  
<term tid="t962_4" lemma="to" pos="TO">  
<target id="w962_4"/> </term>  
<term tid="t962_5" lemma="answer" pos="VB">  
<target id="w962_5"/> </term>  
[...]  
</terms>

</KAF>
Word Alignment

• Ideally the word alignment of testset should be done manually
  • In absence of gold standard, one can use word alignment tools

• To cater for the small testsets (1k), we append each testset to Europarl corpus, and then align them in order to obtain more reliable estimates.

• An advantage of using the TO format is that it allows us to compute union / intersection of alignments produced by different word alignment tools (e.g., GIZA++ and BerkeleyAligner) to improve precision / recall of word alignment.
Word Alignment

Word alignments in TO format for the Italian–English sentence pair:

“È difficile rispondere”

“That is hard to answer”
Kybots

Extract the linguistic phenomena of the source language that are to be evaluated

A Kybot profile can be thought of as a regular expression over elements and attributes in KAF documents.

```
<Kybot id="kybot_v_v">
  <variables>
    <var name="X" type="term" pos="VER*" />
    <var name="Y" type="term" pos="ADJ*" />
    <var name="Z" type="term" pos="VER*" />
  </variables>
  <relations>
    <root span="X" />
    <rel span="Y" pivot="X" direction="following" immediate="true" />
    <rel span="Z" pivot="Y" direction="following" immediate="true" />
  </relations>
</Kybot>
```

Kybot for the linguistic checkpoint “verb_adjective_verb”
N-gram evaluation

• To calculate the score we use a BLEU-style n-gram metric
• For each system-generated translation and reference for a checkpoint
  • Split into a set of n-grams
  • Compute the number of matching n-grams
  • Sum up gains over all n-grams as the score for the checkpoint
• Equation
  • \( R(C) \), recall of a checkpoint \( C \)
  • \( r \), references of the checkpoint

\[
R(C) = \frac{\sum_{r \in R} \sum_{n-gram \in r} \text{match}(n-gram)}{\sum_{r \in R} \sum_{n-gram \in r} \text{count}(n-gram)}
\]

• If the reference of the checkpoint is not consecutive, a wildcard character ("*") is used, which can be matched by any word sequence.
N-gram evaluation

• Consecutive checkpoint

Checkpoint: “È difficile rispondere”
Reference: “that is hard to answer”
Candidate translation: “it is difficult to answer”
Matched n-grams: “is”, “to”, “answer”, “to answer” (4/15)

• Non-consecutive checkpoint

Checkpoint: “È * rispondere”
Reference: “that is * answer”
Candidate translation: “it is difficult to answer”
Matched n-grams: “is”, “answer”, “is * answer” (4/6)
Evaluation

MT systems

- Google Translate
- Bing Translator
- Systran
- FreeTranslation
- CoSyne MT system at Month 12

Language pairs

- DE, IT, NL → EN

Testsets: 1k sentences, news domain
Evaluation

Checkpoints

• Common: n, v, a, r, pre, pro
• DE, IT→EN: frequent 3|4-gram sequences
  • IT→EN
    • n pre-art n
    • det|pre-art n pre|pre-art n
  • DE→EN: pre a n, art a n
• IT→EN: polysemous, n di n
## Evaluation. Results de→en

<table>
<thead>
<tr>
<th>Checkpoints</th>
<th>Instances</th>
<th>Google</th>
<th>Bing</th>
<th>Freetranslation</th>
<th>Systran</th>
<th>CoSyne M12</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>3709</td>
<td>0.6037&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>20.5462&lt;sup&gt;c,e&lt;/sup&gt;</td>
<td>50.5112</td>
<td>30.5384&lt;sup&gt;c,e&lt;/sup&gt;</td>
<td>40.5</td>
</tr>
<tr>
<td>v</td>
<td>1465</td>
<td>0.4836&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>20.4452&lt;sup&gt;c,d,e&lt;/sup&gt;</td>
<td>30.4089</td>
<td>40.4082</td>
<td>50.4054</td>
</tr>
<tr>
<td>a</td>
<td>1004</td>
<td>0.5568&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>20.5229&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>40.4841</td>
<td>50.4781</td>
<td>30.5010</td>
</tr>
<tr>
<td>r</td>
<td>553</td>
<td>0.4232&lt;sup&gt;c&lt;/sup&gt;</td>
<td>30.3858</td>
<td>50.3577</td>
<td>40.3764</td>
<td>20.3989</td>
</tr>
<tr>
<td>pre</td>
<td>1458</td>
<td>0.6058&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
<td>30.5629</td>
<td>50.5355</td>
<td>40.5559</td>
<td>20.5791&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>pro</td>
<td>768</td>
<td>0.5417&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>20.4946&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>50.3616</td>
<td>40.4153&lt;sup&gt;c&lt;/sup&gt;</td>
<td>30.4570&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>appr_art_n</td>
<td>348</td>
<td>0.4212&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
<td>30.3835&lt;sup&gt;c&lt;/sup&gt;</td>
<td>50.3453</td>
<td>40.3825&lt;sup&gt;c&lt;/sup&gt;</td>
<td>20.4110&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>art_adja_n</td>
<td>333</td>
<td>0.4791&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>30.4461&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>50.4022</td>
<td>40.4146</td>
<td>20.4715&lt;sup&gt;c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>BLEU</td>
<td></td>
<td>0.2477&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>20.2294&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>50.1657</td>
<td>40.1752&lt;sup&gt;c&lt;/sup&gt;</td>
<td>30.2052&lt;sup&gt;c,d&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Diagnostic Evaluation with Linguistic Checkpoints
### Evaluation. Results it→en

<table>
<thead>
<tr>
<th>Checkpoints</th>
<th>Instances</th>
<th>Google</th>
<th>Bing</th>
<th>Freetranslation</th>
<th>Systran</th>
<th>CoSyne M12</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>5955</td>
<td>0.74&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.6576&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.5859</td>
<td>0.5947</td>
<td>0.6817&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>v</td>
<td>2959</td>
<td>0.6644&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.5748&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.5119</td>
<td>0.5236</td>
<td>0.6037&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>a</td>
<td>2304</td>
<td>0.7170&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.6376&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.5602</td>
<td>0.5649</td>
<td>0.6651&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>r</td>
<td>837</td>
<td>0.6658&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.5696&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.5124</td>
<td>0.5150</td>
<td>0.5982&lt;sup&gt;c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>pre</td>
<td>3462</td>
<td>0.7369&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.6561&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.6149</td>
<td>0.6198</td>
<td>0.6921&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>pro</td>
<td>981</td>
<td>0.7339&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.6563&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.5776</td>
<td>0.5998</td>
<td>0.6885&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>polysemous</td>
<td>5725</td>
<td>0.7062&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.6231&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.5550</td>
<td>0.5638</td>
<td>0.6574&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>pos_seq3</td>
<td>1075</td>
<td>0.5670&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.4866&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.4071</td>
<td>0.4275</td>
<td>0.5185&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>pos_seq4</td>
<td>195</td>
<td>0.5504&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.4170</td>
<td>0.3929</td>
<td>0.4078</td>
<td>0.4986&lt;sup&gt;b,c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>n_di_n</td>
<td>773</td>
<td>0.5743&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.4991&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.4228</td>
<td>0.4415</td>
<td>0.5270&lt;sup&gt;c,d&lt;/sup&gt;</td>
</tr>
<tr>
<td>BLEU</td>
<td></td>
<td>0.4235&lt;sup&gt;b,c,d,e&lt;/sup&gt;</td>
<td>0.3106&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td>0.1754</td>
<td>0.1840</td>
<td>0.3137&lt;sup&gt;c,d&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

**Diagnostic Evaluation with Linguistic Checkpoints**
## Evaluation. Results nl→en

<table>
<thead>
<tr>
<th>Checkpoints</th>
<th>Instances</th>
<th>Google</th>
<th>Bing</th>
<th>Freetranslation</th>
<th>Systran</th>
<th>CoSyne M12</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>7016</td>
<td>$0.6638_{b,c,d,e}$</td>
<td>$0.6296_{c,d,e}$</td>
<td>$0.5615$</td>
<td>$0.5511$</td>
<td>$0.5615$</td>
</tr>
<tr>
<td>v</td>
<td>2152</td>
<td>$2^0.4503_{c,d,e}$</td>
<td>$0.4508_{c,d,e}$</td>
<td>$0.4261$</td>
<td>$0.4213$</td>
<td>$0.4246$</td>
</tr>
<tr>
<td>a</td>
<td>1992</td>
<td>$0.7019_{b,c,d,e}$</td>
<td>$0.6747_{c,d,e}$</td>
<td>$0.6315^d$</td>
<td>$0.6023$</td>
<td>$0.6481^d$</td>
</tr>
<tr>
<td>r</td>
<td>534</td>
<td>$2^0.4725$</td>
<td>$0.4843$</td>
<td>$0.4510$</td>
<td>$0.4451$</td>
<td>$0.4627$</td>
</tr>
<tr>
<td>pre</td>
<td>3365</td>
<td>$0.6760_{c,d}^a$</td>
<td>$0.7042_{a,c,d,e}$</td>
<td>$0.6148$</td>
<td>$0.6355^c$</td>
<td>$0.6766_{c,d}^a$</td>
</tr>
<tr>
<td>pro</td>
<td>639</td>
<td>$2^0.4764$</td>
<td>$0.4878^c$</td>
<td>$0.4537$</td>
<td>$0.4602$</td>
<td>$0.4439$</td>
</tr>
<tr>
<td>BLEU</td>
<td></td>
<td>$0.3330_{c,d,e}^b$</td>
<td>$0.3347_{c,d,e}$</td>
<td>$0.2456$</td>
<td>$0.2643^c$</td>
<td>$0.3223_{c,d}^e$</td>
</tr>
</tbody>
</table>

Diagnostic Evaluation with Linguistic Checkpoints
Conclusions

• Allows to conduct **fine-grained evaluation**, beyond automatic metrics

• The aim is **not to replace automatic metrics, but to supplement them**

• Knowledge derived from diagnostic evaluation is crucial to
  
  • MT developers, to determine which linguistic phenomena their MT systems are good at dealing with and where they fall behind
  
  • End-users, who might decide to choose a particular MT system over another based on its capability to handle certain linguistic phenomena
Future Work

• Combining different word aligners to improve precision / recall

• Supporting different evaluation metrics

• Developing complex evaluation metric(s)

• Supporting evaluation targets with information up to the level of parsing

• Developing a complete suite of evaluation targets
谢谢 / Thank you for your attention! Questions?

A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints

Sudip Kumar Naskar†‡, Antonio Toral†, Federico Gaspari†, and Andy Way*

NCLT† / CNGL‡ School of Computing
Dublin City University
Dublin 9, Ireland
{snaskar, atoral, fgaspari}@computing.dcu.ie

Applied Language Solutions*
Huddersfield Road
Delph, UK
andy.way@appliedlanguage.com
DE → EN results

<table>
<thead>
<tr>
<th>Translators</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>METEOR-NEXT</th>
<th>TERp</th>
<th>TER</th>
<th>GTM</th>
<th>DCU-LFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>0.2477</td>
<td>0.6358</td>
<td>0.5830</td>
<td>0.4977</td>
<td>0.4000</td>
<td>0.4172</td>
<td>0.4517</td>
<td>0.4899</td>
</tr>
<tr>
<td>Bing</td>
<td>0.2294</td>
<td>0.6362</td>
<td>0.5584</td>
<td>0.4807</td>
<td>0.3600</td>
<td>0.4161</td>
<td>0.4270</td>
<td>0.4570</td>
</tr>
<tr>
<td>Systran</td>
<td>0.1752</td>
<td>0.5447</td>
<td>0.5239</td>
<td>0.4552</td>
<td>0.3216</td>
<td>0.3444</td>
<td>0.4057</td>
<td>0.4133</td>
</tr>
<tr>
<td>Freetranslation</td>
<td>0.1657</td>
<td>0.5212</td>
<td>0.5060</td>
<td>0.4422</td>
<td>0.3100</td>
<td>0.3273</td>
<td>0.3849</td>
<td>0.3957</td>
</tr>
<tr>
<td>CoSyne M12</td>
<td>0.2052</td>
<td>0.5788</td>
<td>0.5470</td>
<td>0.4692</td>
<td>0.2941</td>
<td>0.3700</td>
<td>0.3914</td>
<td>0.4261</td>
</tr>
</tbody>
</table>

Diagnostic Evaluation with Linguistic Checkpoints
IT → EN results

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Bing</th>
<th>Freetranslation</th>
<th>CoSyne M12</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.4235b,c,d,e</td>
<td>0.3106c,d</td>
<td>0.1840d</td>
<td>0.3137c,d</td>
</tr>
<tr>
<td>NIST</td>
<td>0.8579b,c,d,e</td>
<td>0.7517c,d,e</td>
<td>0.5439d</td>
<td>0.7318c,d</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.7017</td>
<td>0.6384</td>
<td>0.5709</td>
<td>0.6565</td>
</tr>
<tr>
<td>METEOR-</td>
<td>0.5942</td>
<td>0.5412</td>
<td>0.4832</td>
<td>0.5545</td>
</tr>
<tr>
<td>NEXT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TERp</td>
<td>0.5600</td>
<td>0.4700</td>
<td>0.3890</td>
<td>0.4946</td>
</tr>
<tr>
<td>TER</td>
<td>0.5599</td>
<td>0.4857</td>
<td>0.3225</td>
<td>0.4679</td>
</tr>
<tr>
<td>GTM</td>
<td>0.6187b,c,d,e</td>
<td>0.5394c,d</td>
<td>0.4596d</td>
<td>0.5475c,d</td>
</tr>
<tr>
<td>DCU-LFG</td>
<td>0.6400</td>
<td>0.5200</td>
<td>0.4244</td>
<td>0.5311</td>
</tr>
</tbody>
</table>

Diagnostic Evaluation with Linguistic Checkpoints
## NL → EN results

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Bing</th>
<th>Systran</th>
<th>Freetranslation</th>
<th>CoSyne M12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BLEU</strong></td>
<td>0.3330</td>
<td>0.3347</td>
<td>0.2643</td>
<td>0.2456</td>
<td>0.3223</td>
</tr>
<tr>
<td><strong>NIST</strong></td>
<td>0.7986</td>
<td>0.7596</td>
<td>0.6830</td>
<td>0.6479</td>
<td>0.7532</td>
</tr>
<tr>
<td><strong>METEOR</strong></td>
<td>0.6633</td>
<td>0.6695</td>
<td>0.6161</td>
<td>0.5964</td>
<td>0.6431</td>
</tr>
<tr>
<td><strong>METEOR-NEXT</strong></td>
<td>0.5583</td>
<td>0.5628</td>
<td>0.5180</td>
<td>0.5032</td>
<td>0.5419</td>
</tr>
<tr>
<td><strong>TERp</strong></td>
<td>0.4987</td>
<td>0.5066</td>
<td>0.4315</td>
<td>0.4123</td>
<td>0.4690</td>
</tr>
<tr>
<td><strong>TER</strong></td>
<td>0.5251</td>
<td>0.4892</td>
<td>0.4424</td>
<td>0.4221</td>
<td>0.5000</td>
</tr>
<tr>
<td><strong>GTM</strong></td>
<td>0.5339</td>
<td>0.5156</td>
<td>0.4761</td>
<td>0.4672</td>
<td>0.4956</td>
</tr>
</tbody>
</table>

### Diagnostic Evaluation with Linguistic Checkpoints

**DCU-LFG**
- 0.5459
- 0.5507
- 0.4661
- 0.4411
- 0.5080
Summary of results

• The three SMT systems receive (much) higher scores than the two RBMT systems for all the 8 evaluation metrics in each of the 3 language pairs

• Overall Google Translate receives the best scores consistently across most of the metrics for all 3 language pairs

• Bing Translator and the CoSyne MT system perform similarly
  ◦ Inferior than Google, but better than Systran and FreeTranslation

• CoSyne good for IT / NL → EN, improvement needed for DE → EN

• Among the RBMT systems, Systran always performs better than FreeTranslation according to all the 8 evaluation metrics for the 3 language pairs