Machine Translation for Multilingual Troubleshooting in the IT Domain: A Comparison of Different Strategies

Sanja Štajner, João Rodrigues, Luís Gomes and António Branco
Department of Informatics, Faculty of Sciences
University of Lisbon, Portugal
Outline

- Problem
- Strategies
- Methodology
- Results
- Conclusions
Problem

- English-Portuguese MT is rarely addressed
- No studies addressing this problem for specific domains
- Domain-specific parallel corpora (EN-PT) are scarce
Strategies

1. Adding out-of-domain corpora
2. Adding in-domain bilingual terminology
3. Adding combination of both (out-of-domain corpora and in-domain bilingual terminology)
Focus

- English to Portuguese MT
- Short sentences (user questions followed by answers from an IT technician)
- Continuous chats
Corpora

1. **EP** – English to Portuguese Europarl (1,960,407 sentence pairs) as the large out-of-domain corpus

2. **IT1** – An in-domain IT corpus with 2,000 sentence pairs (1,000 questions and 1,000 answers) compiled under the QTLeap project (used for training)

3. **IT2** – An in-domain IT corpus with 1,000 sentence pairs (answers only) compiled under the QTLeap project (used for testing)

4. **TERM** – A parallel corpus of IT terminology (unigrams or multiword expressions), which consists of the Microsoft Terminology Collection (13,030 terms) and a small portion of LibreOffice terminology (995 terms).
## Examples

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Source (EN)</th>
<th>Target (PT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TERM</td>
<td>arrow key&lt;br&gt;gatekeeper&lt;br&gt;Planning System Database</td>
<td>tecla de seta&lt;br&gt;controlador de chamadas&lt;br&gt;Base de Dados do Sistema de Planeamento</td>
</tr>
<tr>
<td>IT1</td>
<td>If your disc is not recognized, try changing the USB port.&lt;br&gt;Which antivirus should I keep, MSE or AVG?</td>
<td>Se o disco não está a ser reconhecido, tente trocar de entrada USB.&lt;br&gt;Qual antivrus devo manter, MSE ou AVG?</td>
</tr>
<tr>
<td>IT2</td>
<td>In the Insert menu, select Picture.&lt;br&gt;In the taskbar there is an icon shaped like binoculars, click and type in what you want to search.</td>
<td>No menu inserir selecione Imagem.&lt;br&gt;Na barra de Tarefas há um ícone em forma de binóculos, clique e escreva o que pretende procurar.</td>
</tr>
<tr>
<td>EP</td>
<td>Please rise, then, for this minute’s silence.&lt;br&gt;You have requested a debate on this subject in the course of the next few days, during this part-session.</td>
<td>Convido-os a levantarem-se para um minuto de silêncio.&lt;br&gt;Os senhores manifestaram o desejo de se proceder a um debate sobre o assunto nos próximos dias, durante este período de sessões.</td>
</tr>
</tbody>
</table>
Experiments

MT Systems:
- A hybrid MT system (TectoMT)
- A standard PBSMT system (Moses)

Training datasets:
- IT+TERM (adding terminology)
- IT+EP1 (adding out-of-domain data)
- IT+EP10
- IT+EP10+TERM (adding both)
TectoMT

Tree-to-tree Maximum Entropy Translation Model

TRANSFER

TRANSACTION (EN)

A-tree

Dependencies

NLP tools

Surface string

SYNTHESIS (PT)

A-tree

Surface string

rule-based

rule-based

rule-based

(with the help of NLP tools)
A-tree vs. T-tree

“Try pressing the F11 key.” translated into “Tente carregar na tecla f11.”
Human Evaluation Parameters

Scores:
- Fluency (1 – 4)
- Adequacy (1 – 4)

Error Analysis:
- Orthographic (0 – 2)
- Morphologic (0 – 2)
- Syntactic (0 – 2)
- Semantic (0 – 2)

1 – very bad
2 – bad
3 – good
4 – very good

0 – no errors
1 – one error
2 – two or more errors
## Results (Automatic Evaluation)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Training</th>
<th>Dev.</th>
<th>Test</th>
<th>Results (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EP</td>
<td>TERM</td>
<td>IT1</td>
<td>IT1</td>
</tr>
<tr>
<td>BaselineEP</td>
<td>all</td>
<td>/</td>
<td>/</td>
<td>2,000</td>
</tr>
<tr>
<td>BaselineIT</td>
<td>/</td>
<td>/</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>IT+TERM</td>
<td>/</td>
<td>14,025</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>IT+EP1</td>
<td>1,000</td>
<td>/</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>IT+EP10</td>
<td>10,000</td>
<td>/</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>IT+EP10+TERM</td>
<td>10,000</td>
<td>14,025</td>
<td>2,000</td>
<td>2,000</td>
</tr>
</tbody>
</table>

### TectoMT:
- All above the baselines
- Best approach: both (IT+EP10+TERM)

### PBSMT:
- Above the baselines only those with added terminology (IT+TERM and IT+EP10+TERM)
- Adding a small portion of out-of-domain corpus negatively influences (IT+EP1)
Results (Aspect)

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Sign.</th>
<th>IAA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TectoMT</td>
<td>PBSMT</td>
<td>TectoMT</td>
<td>PBSMT</td>
<td>TectoMT</td>
</tr>
<tr>
<td>Fluency</td>
<td>1.78</td>
<td>1.74</td>
<td>2</td>
<td>1.5</td>
<td>2</td>
</tr>
<tr>
<td>Adequacy</td>
<td>2.28</td>
<td>2.24</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>2.27</td>
<td>2.23</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

- TectoMT achieved significantly higher Adequacy score and Total score
- TectoMT achieved higher Mean and Median value for Fluency (not statistically significant difference)
## Results (Errors)

<table>
<thead>
<tr>
<th>Errors</th>
<th>Mean TectoMT</th>
<th>Mean PBSMT</th>
<th>Median TectoMT</th>
<th>Median PBSMT</th>
<th>Mode TectoMT</th>
<th>Mode PBSMT</th>
<th>Sign.</th>
<th>IAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthographic</td>
<td>1.15</td>
<td>0.95</td>
<td>1.25</td>
<td>1</td>
<td>1.5</td>
<td>1</td>
<td>0.001</td>
<td>0.50</td>
</tr>
<tr>
<td>Morphologic</td>
<td>0.97</td>
<td>0.74</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>0.000</td>
<td>0.54</td>
</tr>
<tr>
<td>Syntactic</td>
<td>1.31</td>
<td>1.26</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>0.045</td>
<td>0.49</td>
</tr>
<tr>
<td>Semantic</td>
<td>1.37</td>
<td>1.50</td>
<td>1.5</td>
<td>1.5</td>
<td>2</td>
<td>2</td>
<td>0.009</td>
<td>0.53</td>
</tr>
</tbody>
</table>

- Number of Orthographic, Morphologic, and Syntactic errors is significantly higher in TectoMT than in PBSMT system.

- Number of Semantic errors is significantly higher in PBSMT than in TectoMT system.
Sentence-wise Comparison

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Scores</th>
<th>Number of errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>TectoMT&gt;PBSMT</td>
<td>47</td>
<td>55</td>
</tr>
<tr>
<td>TectoMT=PBSMT</td>
<td>117</td>
<td>96</td>
</tr>
<tr>
<td>TectoMT&lt;PBSMT</td>
<td>36</td>
<td>49</td>
</tr>
</tbody>
</table>

- Sentences generated by TectoMT represent more fluent and adequate translations, but they also have greater number of errors.

- These results indicate one of the following:
  - Fluency and adequacy cannot be well captured by these types of errors.
  - The errors produced by the TectoMT system are not as severe as those produced by the PBSMT system.
Conclusions

- Adding in-domain bilingual terminology significantly improves the performance of both systems (TectoMT and PBSMT).

- Adding a combination of in-domain bilingual terminology and out-of-domain sentence pairs significantly improves the performance of both systems (TectoMT and PBSMT).

- Adding only some portion of out-of-domain sentence pairs only improves the performance of TectoMT system, while it either impairs or does not significantly change the performance of the PBSMT system.
Limitations

- We used only the basic domain-adaptation technique for the PBSMT system.

- We used no domain-adaptation techniques for the TectoMT system.
Thank you!

Contact:
Sanja.stajner@di.fc.ul.pt
Joao.rodrigues@di.fc.ul.pt
Luis.gomes@di.fc.ul.pt
Antonio.branco@di.fc.ul.pt