Tapta:
A user-driven translation system for patent documents based on domain-aware Statistical Machine Translation

Bruno Pouliquen, Christophe Mazenc, Aldo Iorio
WIPO (World Intellectual Property Organization)
Menu

- Introduction
- Training our SMT
  - English French parallel corpus
  - SMT training
  - Optimization
- Translating GUI
  - Server
  - TAPTA GUI
- Results/Evaluation
  - Automatic evaluation
  - Human evaluation
- Conclusion
Introduction

- WIPO is a UN agency in charge of Intellectual Property
  - Patent Cooperation Treaty => facilitate patent protection in multiple countries
- WIPO receives ~ 150,000 patent applications/year
- Titles and abstracts must be available in English and French
- WIPO has a corpus of 1.8 M en-fr docs (> 8M segments)

- Needs in CAT tools + huge parallel corpus => SMT
- **Translation Assistant** for **Patent Titles and Abstracts**
Our SMT

<table>
<thead>
<tr>
<th>cn</th>
<th>fr</th>
<th>domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>antimicrobial coatings</td>
<td>enrobages antimicrobiens</td>
<td>CHEM</td>
</tr>
<tr>
<td>ternary mixed ethers</td>
<td>éthers mixtes ternaires</td>
<td>CHEM</td>
</tr>
<tr>
<td>submarine</td>
<td>sous – marin</td>
<td>MARI</td>
</tr>
<tr>
<td>automatic translation</td>
<td>traduction automatique</td>
<td>DATA</td>
</tr>
<tr>
<td>automatic translation</td>
<td>translation automatique</td>
<td>BLDG</td>
</tr>
</tbody>
</table>

Parallel segments
Training our SMT: corpus

- Previously translated patent applications (1,800,788 documents)
- Title+abstract
- Additional valuable information:
  - Classification (IPC)
  - Language of publication
  - Quality control passed
- **8’352’768** aligned pairs of segments (high quality)
a pesticide and/or a fertilizer and the like scattered or sprayed onto a golf course \(^\text{(1)}\), together with most of the sprayed water or rain water, are permated into a turf \(^\text{(15A)}\), a thick sand layer \(^\text{(15)}\) having high water permeability, whereby only the rain water filtered by the golf course is discharged to the outside.
SMT Training

- Using open-source Moses
- Training: > 8 Million segments
- Small test set: 102 segments
- Big test set: newly published documents (9521 segments)
“domain-aware” SMT

- Text translated according to the technical domain
- Map IPC code to one of our 32 domains

Domain information as factor, eg:

\[
\text{in\textsuperscript{|U}} \text{ the\textsuperscript{|U}} \text{ robotic\textsuperscript{|1}} \text{ prosthesis\textsuperscript{|1}} \text{ alignment\textsuperscript{|1}} \text{ device\textsuperscript{|U}} \text{ provides\textsuperscript{|U}} \text{ automatic\textsuperscript{|1}} \text{ translation\textsuperscript{|1}} \text{ in\textsuperscript{|U}} \text{ two\textsuperscript{|U}} \text{ axes\textsuperscript{|1}} \text{ a\textsuperscript{|U}} \text{ chinese\textsuperscript{|2}} \text{ to\textsuperscript{|U}} \text{ english\textsuperscript{|2}} \text{ automatic\textsuperscript{|2}} \text{ translation\textsuperscript{|2}} \text{ method\textsuperscript{|U}}
\]
SMT… more…

- Language model: irstlm (5 grams), kenlm, binarization
- Pruning of the phrase table, binarized (102M entries x 2)
- Improve the quality:
  - Get only segments out of QC-passed documents
  - Add segments that were originally in French
  - Create a sub-model (phrase tables)
  - Merge this sub-model with original (75-25%)
Graphical User Interface

Live demo???
# Evaluation: automatic

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Speed Seconds (#docs)</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>542 (51)</td>
<td>40.11</td>
</tr>
<tr>
<td>Pruned</td>
<td>164 (51)</td>
<td>41.11</td>
</tr>
<tr>
<td>Pruned+d=0.15+s=10</td>
<td>22 (51)</td>
<td>43.30</td>
</tr>
<tr>
<td>Pruned+BigLm</td>
<td>196 (51)</td>
<td>51.71</td>
</tr>
<tr>
<td>Pruned+BigLm w/oDomain</td>
<td>173 (51)</td>
<td>50.91</td>
</tr>
<tr>
<td>Only Qc</td>
<td>N/a (51)</td>
<td>28.29</td>
</tr>
<tr>
<td>Google translate</td>
<td>N/a (51)</td>
<td>34.09</td>
</tr>
<tr>
<td>Bing translator</td>
<td>N/a (51)</td>
<td>27.17</td>
</tr>
<tr>
<td>Pruned+BigLm</td>
<td>27140 (1390)</td>
<td>45.06</td>
</tr>
<tr>
<td>Pruned+BigLm w/oDomain</td>
<td>6934 (1390)</td>
<td>44.99</td>
</tr>
<tr>
<td>Pruned+BigLm_Qc_d=0.15 s=10</td>
<td>6378 (1390)</td>
<td>45.07</td>
</tr>
</tbody>
</table>

**Reasonable speed: 5 sec/doc (without parallelization)**

**Real test: translate latest published documents**
Evaluation: human

- 12 testers
  - Good French and English skills
  - Not translators
- Coached by 2 professional translator-revisers
- 516 translations produced during 13 days

<table>
<thead>
<tr>
<th></th>
<th>Translations produced</th>
<th>😊 Publishable</th>
<th>😞 Non Publishable</th>
</tr>
</thead>
<tbody>
<tr>
<td>With coaching (10 days)</td>
<td>403</td>
<td>243</td>
<td>60.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>160</td>
<td>39.7%</td>
</tr>
<tr>
<td>Without coaching (3 days)</td>
<td>113</td>
<td>56</td>
<td>49.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>57</td>
<td>50.4%</td>
</tr>
</tbody>
</table>

~ half of translation produced by non translators passed a high quality QC
Second evaluation: compare with outsourcing agencies

<table>
<thead>
<tr>
<th>Translations</th>
<th>translations QCed</th>
<th>Tapta</th>
<th>Agency</th>
<th>Same quality</th>
<th>Tapta better</th>
<th>Agency better</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>NP</td>
<td>P</td>
<td>NP</td>
<td></td>
</tr>
<tr>
<td>With coaching</td>
<td>388</td>
<td>350</td>
<td>38</td>
<td>344</td>
<td>44</td>
<td>111</td>
</tr>
<tr>
<td>Without coaching</td>
<td>112</td>
<td>103</td>
<td>9</td>
<td>98</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td>453</td>
<td>47</td>
<td>442</td>
<td>58</td>
<td>134</td>
</tr>
<tr>
<td>%</td>
<td>100</td>
<td>90.6</td>
<td>9.4</td>
<td>88.4</td>
<td>11.6</td>
<td>26.8</td>
</tr>
</tbody>
</table>
Conclusion

- SMT with reasonable results
- User-driven translation was successful
- Tapta was not adopted by WIPO professional translators
- But:
  - Adding Tapta in the workflow is investigated
  - Was judged as very valuable for non-translators
  - Accelerated training aid
Future work

- Other language pairs: Chinese, Korean, Japanese ...
- Triangulation
- Incremental training

Web Tapta: http://www.wipo.int/patentscope/translate/