Machine Translation and Multilingual Search for the Patent Domain: PLuTO

Mihai Lupu
IRF / ESTeam

Information Retrieval Facility Symposium
June 7th, 2011
Vienna, AT

Funded under the EU ICT Policy Support Programme
Overview

Introduction
  Research on Patent Data
  What is PLuTO?

PLuTO - Patent Language Translations Online
  Objectives
  Consortium
  Components
  Where we are now

Demo
Research on patent data

- Observed increased interest
  - ACM digital library statistics
  - Springer publisher statistics
Research on patent data

- Stimulate interest
  - Funding (as described)
  - Providing data
    - E.g. MAREC
  - Organizing Challenges
    - CLEF-IP
    - TREC-CHEM
  - Organizing fora
    - Patent Information Retrieval Workshop

- Transition from Research to Industry-ready systems
What is PLuTO?

- **Patent Language Translations Online**
  - Industry-Academia partnership

- 3 year funded project - April 2010 → April 2013
  - For a total cost of up to 2.2 million €

- 50% funded
  - Aim: bring state-of-the-art from academia to commercialization level
  - core deliverables focussing on dissemination and exploitation
A number of goals exist in order to satisfy the technological requirements of the project and set the foundations for a viable commercial entity

- Development on an online solution for patent search and translation
  - English, French, German, Spanish, Portuguese, ...

- Provide multilingual access to online digital patent libraries

- Community based web service/collaborative space
  - patent searches

- Disseminate the fruits of the project, build awareness and engage with potential customers
PLuTO Consortium

- Dublin City University, Ireland
  - Machine Translation
- ESTeam, Sweden
  - Translation Memory / Search
- Cross Language, Belgium
  - Evaluation (linguistic)
- Dutch Patent User Group, the Netherlands
  - Evaluation (usability/appropriateness)
PLuTO Components

ESTeam Translator

Comprehensive translation software environment including translation memory (TM).

MaTrEx (CNGL, DCU)

Hybrid phrase-based, example-based and hierarchical phrase-based system approaches. Proven performance in recent MT Evaluation campaigns.

Lucene/SOLR Indexing Engine

State-of-the-art open-source platform for Information Retrieval, including a variety of analyzers and stemmers in various languages.

Patent Data Collections (IRF, WON and others, inc. EPO)

PLuTO integrates existing components and creates a web interface on top of an underlying multilingual database.
PLuTO Components

Statistical Machine Translation

Web interface

Search Engine, data files

Translation Service, TMs
PLuTO - where we are now

1st prototype

- Web interface for search and translation
- Search engine
  - Index 19’386’697 patent files
    - 1978-june 2008
    - EPO, JPO, USPTO, WIPO
  - EN, FR, DE, RU, others
- Machine Translation
  - Statistical MT & Translation Memories
    - EN <-> FR
    - EN <-> PT
    - FR <-> PT (via TM)
Initial Feedback from Users in Year 1

- **Patent Search** is a crowded market with established players offering value-add feature-laden products. Patent collection size/coverage is critical.

- **(Automated) Patent Translation** is an emerging fragmented market with several offerings of disparate quality, pricing models, etc.

- There is an opportunity for a provider with good quality translation (tailored to task, domain). Customers are used to paying for quality results.
Patent Translation: Competitive Landscape

- Google
  - Search/Access all the world’s information - PLuTO focused on patents only. Use patent domain/style/etc. info to improve quality of translation. [Plus our evaluation shows us better than Google]
    - Similarly... Google does search - but patent search is different.

- Moses - open source community
  - Steep learning curve of Moses, adaptation, tuning, etc.
  - [Other projects, Euromatrix+ etc. working on Moses also]
  - MaTrEx MT @ DCU based around Moses ++ improvements

- Patent Offices/Organisations - Korea, Japan, EPO, WIPO
  - Fragmented providers, disparate quality levels, variable pricing models.
Translation Evaluation

- **Automatic evaluation**
  - 8000 sentences in total (1000 per IPC categories A-H)
  - BLEU/METEOR
  - Score per category + overall score

- **Human evaluation**
  - 800 sentences in total (100 per IPC categories A-H)
  - Adequacy evaluation: transfer of meaning PLuTO translations
  - Ranking evaluation: comparison PLuTO vs Google and Systran
  - Overall score only

- 3 informants per language direction
  - 3 evaluators for en->pt, 3 for pt->en, 3 for en->fr, and 3 for fr->en
  - Profile: experienced patent translators familiar with machine translation
## Automatic Evaluation

<table>
<thead>
<tr>
<th>French to English</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td>42.52/59.65</td>
<td>56.92/67.44</td>
<td>28.90/53.67</td>
</tr>
<tr>
<td>A (Human necessities)</td>
<td>43.60/60.58</td>
<td>58.35/68.22</td>
<td>28.05/53.46</td>
</tr>
<tr>
<td>B (Performing Operations)</td>
<td>42.29/59.84</td>
<td>55.03/66.95</td>
<td>30.45/54.53</td>
</tr>
<tr>
<td>C (Chemistry)</td>
<td>46.66/61.81</td>
<td>62.01/70.03</td>
<td>29.92/54.44</td>
</tr>
<tr>
<td>D (Textiles)</td>
<td>42.53/59.35</td>
<td>56.51/67.03</td>
<td>24.49/53.54</td>
</tr>
<tr>
<td>E (Fixed constructions)</td>
<td>40.27/57.29</td>
<td>53.85/64.73</td>
<td>30.12/52.99</td>
</tr>
<tr>
<td>F (Mechanical engineering)</td>
<td>43.28/60.36</td>
<td>57.21/67.77</td>
<td>31.28/55.35</td>
</tr>
<tr>
<td>G (Physics)</td>
<td>40.74/59.51</td>
<td>56.21/67.90</td>
<td>25.55/53.11</td>
</tr>
<tr>
<td>H (Electricity)</td>
<td>41.36/58.89</td>
<td>56.32/67.53</td>
<td>25.89/51.91</td>
</tr>
</tbody>
</table>
## Automatic Evaluation

<table>
<thead>
<tr>
<th>French to English</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>42.52/59.65</td>
<td>56.92/67.44</td>
<td>28.90/53.67</td>
</tr>
</tbody>
</table>

### A (Human necessities)

<table>
<thead>
<tr>
<th>English to Portuguese</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>18.64 / 22.12</td>
<td>37.84 / 40.94</td>
<td>15.36 / 18.79</td>
</tr>
</tbody>
</table>

### B (Performing Operations)

<table>
<thead>
<tr>
<th>English to Portuguese</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>17.71 / 20.96</td>
<td>32.42 / 35.29</td>
<td>13.42 / 16.40</td>
</tr>
</tbody>
</table>

### C (Chemistry)

<table>
<thead>
<tr>
<th>English to Portuguese</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>14.54 / 16.66</td>
<td>26.26 / 28.78</td>
<td>11.1 / 13.09</td>
</tr>
</tbody>
</table>

### D (Textiles)

<table>
<thead>
<tr>
<th>English to Portuguese</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>20.24 / 24.10</td>
<td>39.42 / 42.67</td>
<td>16.8 / 20.58</td>
</tr>
</tbody>
</table>

### E (Fixed constructions)

<table>
<thead>
<tr>
<th>English to Portuguese</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>18.5 / 21.87</td>
<td>38.74 / 41.73</td>
<td>16.21 / 19.59</td>
</tr>
</tbody>
</table>

### F (Mechanical engineering)

<table>
<thead>
<tr>
<th>English to Portuguese</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>19.25 / 22.71</td>
<td>42.21 / 45.21</td>
<td>16.7 / 20.20</td>
</tr>
</tbody>
</table>

### G (Physics)

<table>
<thead>
<tr>
<th>English to Portuguese</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>20.6 / 24.38</td>
<td>44.51 / 47.79</td>
<td>16.41 / 20.12</td>
</tr>
</tbody>
</table>

### H (Electricity)

<table>
<thead>
<tr>
<th>English to Portuguese</th>
<th>Google</th>
<th>Pluto</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>20.17 / 24.37</td>
<td>40.40 / 43.87</td>
<td>17.56 / 21.60</td>
</tr>
</tbody>
</table>
Human evaluation

Initial evaluation results for translation

French to English

<table>
<thead>
<tr>
<th>System</th>
<th>Rank 1</th>
<th>Rank 2</th>
<th>Rank 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>41%</td>
<td>45%</td>
<td>14%</td>
</tr>
<tr>
<td>PLuTO</td>
<td>67%</td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td>Systran</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Source:** A obtenção das nanofibras é feita através da técnica de electrofição descrita na literatura...

**Pluto:** Obtaining the nanofibers is made through the electrofição technique described in the literature...

**Google:** The achievement is made of nanofibres through the technique described in the literature and electrofição...

---

**Source:** Por exemplo, para as estruturas de média importância, uma cadeia de reunião de tabuleiros é composta pelos seguintes postos:

**Pluto:** For example, for the medium importance structures, a chain of meeting trays is composed of the following compounds:

**Google:** For example, for structures of medium importance, a chain of boards meeting consists of the following posts:
Competitive Positioning: Summary

- PLuTO developing focused, optimised solution for patent translation, leveraging unique features of patent collections, texts.

- Builds on over 10 years of research in MT at DCU, improving upon Moses core, while continually leveraging Moses improvements.
  - Engines evaluated independently by Cross Language.

- Unique integration with advanced TM tools from ESTeam for tailored, high-quality and/or human post-edited translations.
Future Plans - *LISTEN* to customers/users

- Patent search is a crowded space with many feature-rich products. Will be a tough market to crack.

- Patent MT is ripe for exploitation. PLuTO has advanced state-of-the-art and MT/TM integration offers differentiation - worth focusing here.

- Focus on language selection - market driven. Huge demand for Asian language <> English MT in patent space.
Proposals to respond to customers’ needs

- **Un-tether MT (from search system) - Browser Plug-in / App.** Works across web with variety of patent search systems/sources.

- **Continue TM/MT integration to drive quality and reduce overall time/cost of human-quality (post-edit) translation.**

- **Market-focused language selection - to include Asian languages.**
Conclusions

PLuTO - a system for Patent Machine Translation

- with Patents and only for Patents
Thank you!

PLuTO website

http://www.pluto-patenttranslation.eu