MT for Online Patent Translation

John Tinsley
Centre for Next Generation Localisation
Dublin City University

AMTA 2010, Denver, CO., 03/11/2010

Funded under the EU ICT Policy Support Programme
Overview

Background
- What is PLUTO?
- Motivation
- Objectives
- Consortium

Machine Translation in PLUTO
- MaTrEx MT System
- Patent Translation

Use Case: Online MT
- Domain Adaptation
- Model Management
Background

- What is PLUTO?
- Motivation
- Objectives
- Consortium
What is PLuTO?

- Patent Language Translations Online
  - Industry-Academia partnership

- 3 year EU funded project under the ICT-PSP scheme
  - Develop inclusive information society
  - Strengthen market for ICT products and services
  - Theme: MT for the multilingual web

- 50% funded commercialisation project
  - core deliverables focussing on dissemination and exploitation
Motivation

- EU has always been a big proponent of HLT, particularly MT

- Committed to the provision of multilingual access to intellectual property information and language diversification
  - This fosters innovation and growth

- Requirements? Translation of back catalogues and services for translation of new IP

- Big factor is the political shift towards a single European patent
PLuTO Objectives

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
  - prior art 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
- Information Retrieval Facility, Austria
  - Search
- Cross Language, Belgium
  - Evaluation (linguistic)
- Dutch Patent User Group, the Netherlands
  - Evaluation (usability/appropriateness)

Funded under the EU ICT Policy Support Programme
Machine Translation in PLuTO

- MaTrEx MT Engine
- Translation Process
- Deployment
- Patent Translation
MT in PLUTO is performed using the data-driven MaTrEx engine developed at DCU

Advantages
- Rapid deployment
- Iterative improvement via offline training

Drawbacks
- Need for training data
- Resource intensive
Translation Process

Before translation
- Cleaning data, tokenisation/casing formatting, job scheduling (task farming)

After translation
- Document reassembly, de-tokenisation/re-casing

Deployment of Service
- Web service (existing) vs. local install
- MT web service:
  - Backend to search functionality
  - Direct translation via API
Patent Translation

Patents present a unique challenge when it comes to MT due to the nature of the language found in documents.

Documents are structured into three sections:
- Abstract - general language
- Claims - legalese
- Descriptions - technical terms

This challenge is exacerbated by the fact that the technical details in a patent can relate to almost anything:
- a search algorithm for MT
- a promotional philosophy for VOD interactive displays
- a rope-walking monkey toy
WIPO - the World Intellectual Property Organisation - has created the International Patent Classification (IPC) system which “provides for a hierarchical system...for the classification of patents...according to the different areas of technology to which they pertain”

Top-level of the IPC:

A. Human Necessities
B. Performing Operations; Transporting
C. Chemistry; Metallurgy
D. Textiles; Papers
E. Fixed Constructions
F. Mechanical Engineering; Lighting; Heating; Weapons
G. Physics
H. Electricity
IPC Codes

- IPC system gets very granular...

“Rope-Walking Monkey Toy”

IPC code: A63H13/12

Breakdown

A - Human necessity (health; life-saving; amusement)

63 - Sports; games; amusements

H - Toys e.g. dolls

13 - with moving parts

12 - acrobatic figures
We have identified two main challenges as relates to the online patent translation service provided by PLUTO:

**Domain Adaptation**
- What level of granularity in IPC?
- Separate systems for abstracts, claims and descriptions?
- Translation model? Language model? Both?

**Model Management**
- 3 factors to consider: translation quality, translation speed, required computational resources
- Find the right balance
We describe experiments we carried out to determine the optimal system configuration for an online translation service address the two aforementioned challenges.

Details on the service:
- European Patent Office’s ‘Espacenet’ service
- English <--> Portuguese translation
- 10,000 translation requests per month
Use Case: domain adaptation

Two decisions made *a priori*
- Only consider the 8 top-level IPC sub-domains
- Do not treat document sections differently

Distribution of data across IPC codes

<table>
<thead>
<tr>
<th>Percentage</th>
<th>IPC Code</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>32%</td>
<td>A.</td>
<td>Human Necessities</td>
</tr>
<tr>
<td>9%</td>
<td>B.</td>
<td>Performing Operations; Transporting</td>
</tr>
<tr>
<td>44%</td>
<td>C.</td>
<td>Chemistry; Metallurgy</td>
</tr>
<tr>
<td>2%</td>
<td>D.</td>
<td>Textiles; Papers</td>
</tr>
<tr>
<td>1%</td>
<td>E.</td>
<td>Fixed Constructions</td>
</tr>
<tr>
<td>2%</td>
<td>F.</td>
<td>Mechanical Engineering; Lighting; Heating; Weapons</td>
</tr>
<tr>
<td>6%</td>
<td>G.</td>
<td>Physics</td>
</tr>
<tr>
<td>4%</td>
<td>H.</td>
<td>Electricity</td>
</tr>
</tbody>
</table>
Use Case: domain adaptation

Two decisions made *a priori*

- Only consider the 8 top-level IPC sub-domains
- Do not treat document sections differently

Distribution of data across IPC codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Human Necessities</td>
<td>32%</td>
</tr>
<tr>
<td>B</td>
<td>Performing Operations; Transporting</td>
<td>9%</td>
</tr>
<tr>
<td>C</td>
<td>Chemistry; Metallurgy</td>
<td>44%</td>
</tr>
<tr>
<td>D</td>
<td>Textiles; Papers</td>
<td>2%</td>
</tr>
<tr>
<td>E</td>
<td>Fixed Constructions</td>
<td>1%</td>
</tr>
<tr>
<td>F</td>
<td>Mechanical Engineering; Lighting; Heating; Weapons</td>
<td>2%</td>
</tr>
<tr>
<td>G</td>
<td>Physics</td>
<td>6%</td>
</tr>
<tr>
<td>H</td>
<td>Electricity</td>
<td>4%</td>
</tr>
</tbody>
</table>
Use Case: domain adaptation

When translating a document in a given domain, we want to establish whether it is better to train an MT system using only patent data in that domain or whether we should use more of the data at our disposal.

Testsets
For each of the 3 domains - B, C and E - we have a test set comprising 1,000 sentences.

MT Systems
We run each test set through 4 MT systems:
- In-domain translation model (TM) and language model (LM)
- In-domain TM and general LM
- General TM and in-domain LM
- General TM and general LM

E.g. for Testset B, ‘in-domain’ means only data from B. ‘General’ means all data from B, C and E.
### Use Case: domain adaptation

<table>
<thead>
<tr>
<th>Testset B</th>
<th>T-Table</th>
<th>L-model</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td>In-domain</td>
<td>0.4845</td>
<td>74.13</td>
<td></td>
</tr>
<tr>
<td>In-domain</td>
<td>General</td>
<td>0.5224</td>
<td>75.58</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>In-domain</td>
<td>0.5281</td>
<td>75.76</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>General</td>
<td>0.5495</td>
<td>76.53</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testset C</th>
<th>T-Table</th>
<th>L-model</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td>In-domain</td>
<td>0.4544</td>
<td>67.58</td>
<td></td>
</tr>
<tr>
<td>In-domain</td>
<td>General</td>
<td>0.4563</td>
<td>67.85</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>In-domain</td>
<td>0.5998</td>
<td>80.74</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>General</td>
<td>0.5998</td>
<td>80.74</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testset E</th>
<th>T-Table</th>
<th>L-model</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td>In-domain</td>
<td>0.4708</td>
<td>73.21</td>
<td></td>
</tr>
<tr>
<td>In-domain</td>
<td>General</td>
<td>0.5171</td>
<td>74.71</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>In-domain</td>
<td>0.5401</td>
<td>76.43</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>General</td>
<td>0.5679</td>
<td>77.44</td>
<td></td>
</tr>
</tbody>
</table>
Use Case: domain adaptation

<table>
<thead>
<tr>
<th>Testset B</th>
<th>T-Table</th>
<th>L-model</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td>In-domain</td>
<td>0.4845</td>
<td>74.13</td>
<td></td>
</tr>
<tr>
<td>In-domain</td>
<td>General</td>
<td>0.5224</td>
<td>75.58</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>In-domain</td>
<td>0.5281</td>
<td>75.76</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>General</td>
<td>0.5495</td>
<td>76.53</td>
<td></td>
</tr>
</tbody>
</table>

Observations

- a general LM improves a lot: sufficient natural language despite technical nature of data in ‘C’
- a general TM also improves a lot

Is there the case for filtering the data from C when building a general set?
Use Case: domain adaptation

<table>
<thead>
<tr>
<th>Testset C</th>
<th>T-Table</th>
<th>L-model</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td>In-domain</td>
<td></td>
<td>0.4544</td>
<td>67.58</td>
</tr>
<tr>
<td>In-domain</td>
<td>General</td>
<td></td>
<td>0.4563</td>
<td>67.85</td>
</tr>
<tr>
<td>General</td>
<td>In-domain</td>
<td></td>
<td>0.5971</td>
<td>80.57</td>
</tr>
<tr>
<td>General</td>
<td>General</td>
<td></td>
<td>0.5998</td>
<td>80.74</td>
</tr>
</tbody>
</table>

Observations
- a general LM does not help; ‘C’ data is so specific, general language from the other do not help with fluency
- a general TM gives huge improvements

The more technical the data, the less useful a general LM is?
Use Case: domain adaptation

Observations
- in all instances, the addition of general domain data improves translation
- this can be attributed to the relatively small size of the training set for domain ‘E’ (1% of total vs. 9% for ‘B’ and 44% for ‘C’)

Obviously the less data available for a particular domain, the more beneficial it will be to add data from other domains. Should data be profiled to determine which domains complement each other better?
Use Case: model management

- As we mentioned, while having a lot of data is beneficial in terms of building quality MT systems, exploiting it efficiently is a challenge.

- 3 key factors: translation time, translation quality, required resources.

- Finding a compromise depends on the requirements of the task at hand.
  - How much weight is given to each?

We build a multitude of systems tweaking various parameters:

- Language model order
- Maximum phrase length
- Phrase table loading method
- Beam/stack size during decoding
Use Case: model management

- Ideally, we would like a small bubble in the top-right corner.
- It is clearly not a straightforward task choosing which configuration to use.
- This will always be dependent on the task at hand.
That is where PLUTO is, as of now

- Live MT system for EN–PT patent translation
- Investigated to some extent the question of domain adaptability
  - Find more concrete answers to those questions
- Issue of model management will always be dictated by the task
The future for PLuTO

- Fully integrated search and translation tool (first prototype due early 2011)
- Translation will feature integrated MT and translation memory
- Web application allowing for collaborative prior art searching

Enterprise PLUTO - Watch this space!
Thank you!

- PLUTO website
  http://www.pluto-patenttranslation.eu

- PLUTO @ Espacenet (En–Pt)
  http://www.espacenet.com