

# 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



## Background

- What is PLUTO?
- Motivation
- Objectives
- Consortium

## Machine Translation in PLUTO

- MaTrEx MT System
- Patent Translation

## Use Case: Online MT

- Domain Adaptation
- Model Management

- 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





Dublin City University, Ireland

- Machine Translation



ESTeam, Sweden

- Translation Memory



**INFORMATION  
RETRIEVAL  
FACILITY**

Information Retrieval Facility, Austria

- Search



Cross Language, Belgium

- Evaluation (linguistic)



Dutch Patent User Group, the Netherlands

- Evaluation (usability/appropriateness)

# Machine Translation in P LuTO

- 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

### Method for producing a fresh cold coffee drink and a corresponding coffee machine

**Bibliographic data** | Description | Claims | Mosaics | Original document | INPADOC legal status

**Publication number:** EP2238876 (A2)  
**Publication date:** 2010-10-13  
**Inventor(s):** BUCHHOLZ BERND DR [DE]; SCHMALKUCHE JENS [DE]; TINELNOT PETER [DE] +  
**Applicant(s):** MELITTA SYSTEMSERVICE GMBH & C [DE] +  
**Classification:**  
- international: A47J31/00; A47J31/057; A47J31/24; A47J31/00; A47J31/04; A47J31/24  
- European:  
**Application number:** EP20100155806 20100308  
**Priority number(s):** DE200910016506 20090408

[View document in the European Register](#)  [Report a data error here](#)

Abstract of EP 2238876 (A2) [Translate this text](#)

Ein Verfahren zum Erzeugen eines frischen Kaffeegetränks aus einer vorgegebenen Menge Kaffeebohnen, mit folgenden Verfahrensschritten: (S1) Zuführen der vorgegebenen Menge Kaffeebohnen in eine Brühkammer (3); (S2) Erstellen eines Kaffeekonzentrats aus der zugeführten vorgegebenen Menge Kaffeebohnen mittels eines Brühvorgangs in der Brühkammer (3), wobei das so erstellte Kaffeekonzentrat ein vorgebares Vielfaches (VK) einer Konzentration (KN) einer normalen Trinkqualität für Filterkaffee mit ca. 1,2 bis 1,4 % Trockensubstanzanteil aufweist; (S3) Erzeugen eines frischen Kaffeegetränks durch Vermischen des so erstellten Kaffeekonzentrats mit einer Wassermenge, welche einem vorgebaren Vielfachen (VM) der Menge des erstellten frischen Kaffeekonzentrats entspricht.

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*"

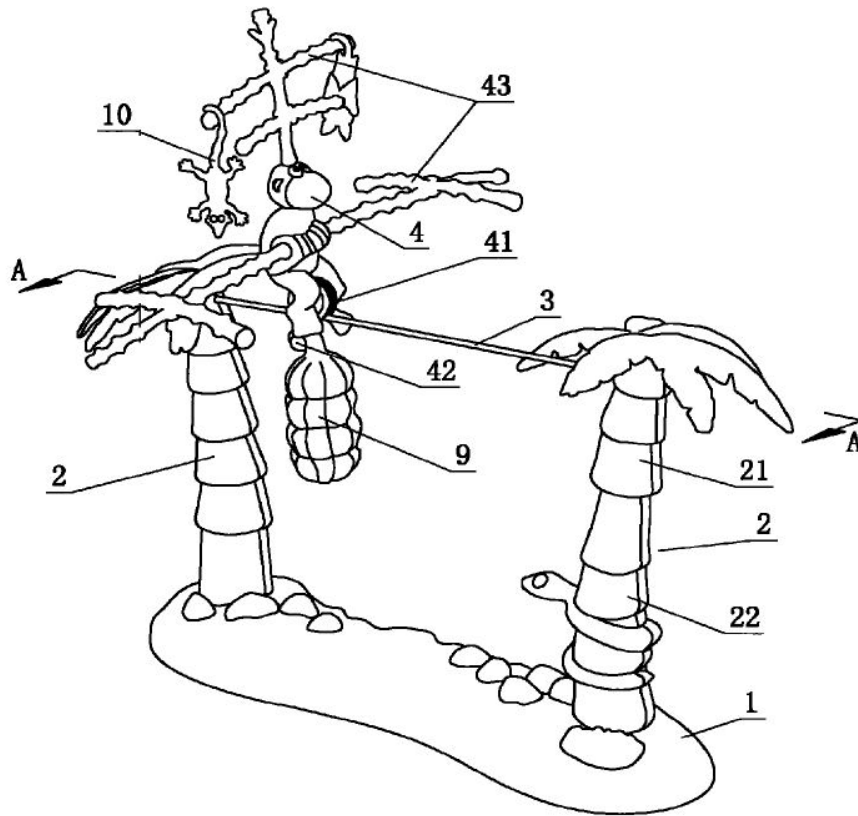


WORLD  
INTELLECTUAL  
PROPERTY  
ORGANIZATION

## 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 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



# Use Case: EPO

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

32%	A. Human Necessities
9%	B. Performing Operations; Transporting
44%	C. Chemistry; Metallurgy
2%	D. Textiles; Papers
1%	E. Fixed Constructions
2%	F. Mechanical Engineering; Lighting; Heating; Weapons
6%	G. Physics
4%	H. Electricity

# 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

32% A. Human Necessities

9% B. Performing Operations; Transporting

44% C. Chemistry; Metallurgy

2% D. Textiles; Papers

1% E. Fixed Constructions

2% F. Mechanical Engineering; Lighting; Heating; Weapons

6% G. Physics

4% H. Electricity

# 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

Testset B

T-Table	L-model	BLEU	METEOR
In-domain	In-domain	0.4845	74.13
In-domain	General	0.5224	75.58
General	In-domain	0.5281	75.76
<i>General</i>	<i>General</i>	<i>0.5495</i>	<i>76.53</i>

Testset E

T-Table	L-model	BLEU	METEOR
In-domain	In-domain	0.4708	73.21
In-domain	General	0.5171	74.71
General	In-domain	0.5401	76.43
<i>General</i>	<i>General</i>	<i>0.5679</i>	<i>77.44</i>

Testset C

T-Table	L-model	BLEU	METEOR
In-domain	In-domain	0.4544	67.58
In-domain	General	0.4563	67.85
General	In-domain	0.5971	80.57
<i>General</i>	<i>General</i>	<i>0.5998</i>	<i>80.74</i>

# Use Case: domain adaptation

Testset B			
T-Table	L-model	BLEU	METEOR
In-domain	In-domain	0.4845	74.13
In-domain	General	0.5224	75.58
General	In-domain	0.5281	75.76
<i>General</i>	<i>General</i>	<i>0.5495</i>	<i>76.53</i>

## 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

Testset C			
T-Table	L-model	BLEU	METEOR
In-domain	In-domain	0.4544	67.58
In-domain	General	0.4563	67.85
General	In-domain	0.5971	80.57
<i>General</i>	<i>General</i>	<i>0.5998</i>	<i>80.74</i>

## 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

Testset E			
T-Table	L-model	BLEU	METEOR
In-domain	In-domain	0.4708	73.21
In-domain	General	0.5171	74.71
General	In-domain	0.5401	76.43
<i>General</i>	<i>General</i>	<i>0.5679</i>	<i>77.44</i>

## 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?

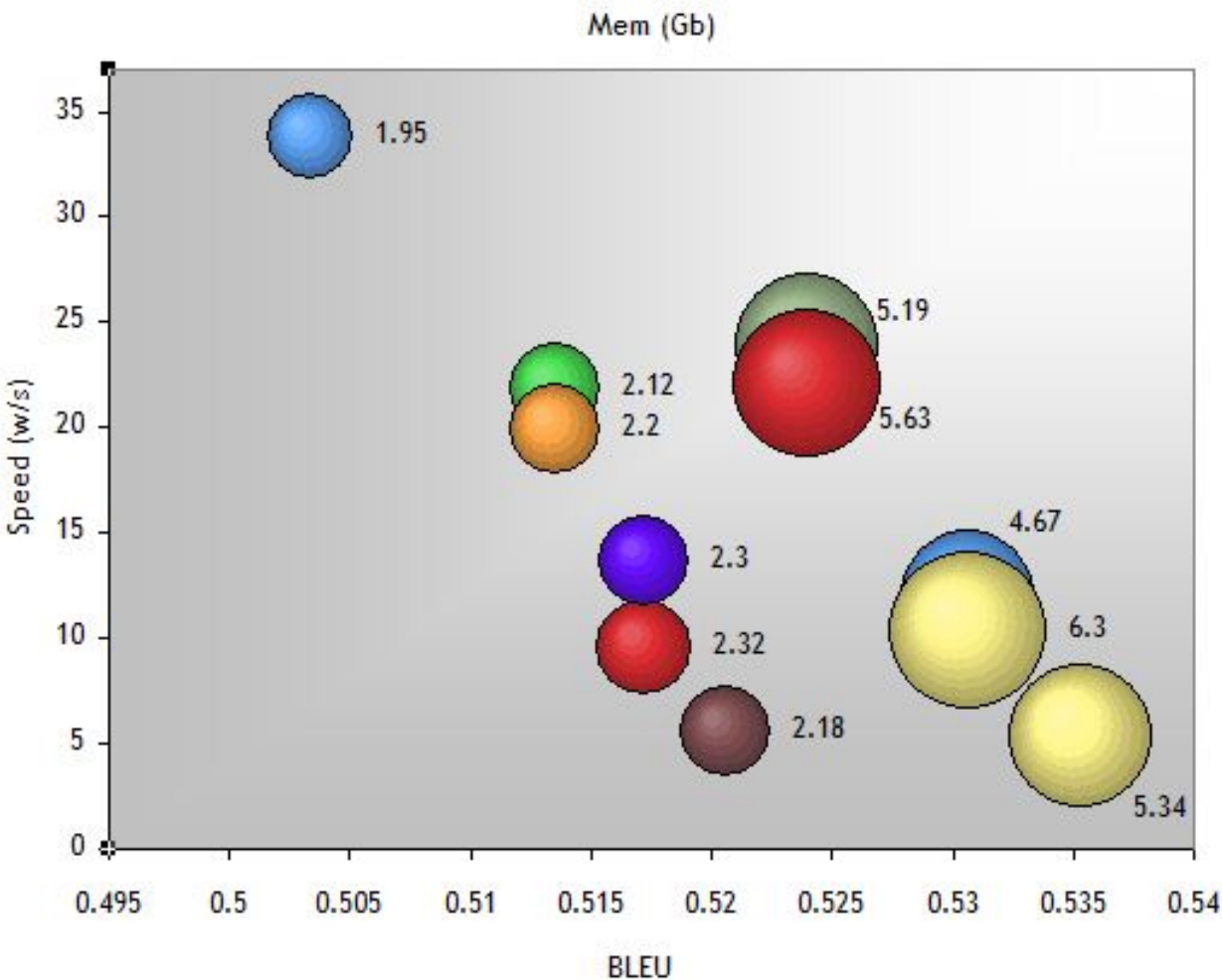
# Use Case: model management

- 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

# “State of the Union”

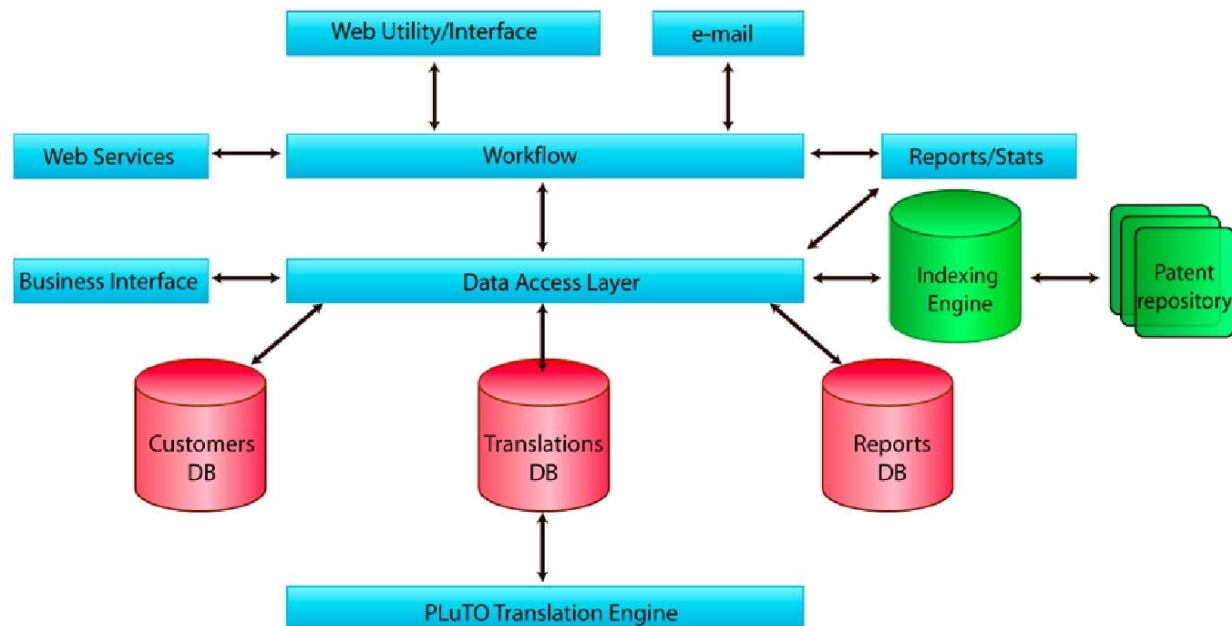
That is where PLUTO is, as of now

- 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>