PLuTo

Patent Language Translations Online

Publishable Summary

Six Month Progress Report

November 2010

The PLUTO project has received funding from the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement n° 250430

Document Status

Project	**PLUTO: CIP ICT-PSP Grant # 250430**
Document version	2.0 FINAL
Primary Author	Dr. John Tinsley, DCU
Review	Dr. Páraic Sheridan, Prof. Andy Way
Status	**Public**

TABLE OF CONTENTS

1. Executive Summary

The PLUTO project (Patent Language Translations Online) aims to provide a rapid solution for the online retrieval and translation of patent documents through the integration of a number of existing state-of-the-art components provided by the partners in our dynamic industry-academia consortium.

It is intended that the provision of this service will have a significant positive impact on the intellectual property (IP) related activities of a number of parties; individual inventors and Small- and Medium-sized Enterprises (SMEs) with potentially limited resources, as well as multinational corporations and other entities, will be afforded the ability to carry out prior art searches in a much more effective and efficient manner through a reduction in time and translations costs when using PLUTO services.

Additionally, PLUTO aims to advance the state-of-the-art in domain focused machine translation (MT) through its integration with translation memories (TM). A number of novel approaches to cross-lingual information retrieval will also be employed during the construction of a multilingual searchable database of patents.

Since the project kick-off meeting in May 2010, PLUTO has seen rapid progress toward its intended goals due to a combination of consortium partners' existing working relationships, the strength of products and systems upon which the project builds, and close alignment with pilot initiatives for Machine Translation at the European Patent Office.

All internal milestones of the project have been met successfully, with data being provided for TM and MT development, documented integration contracts detailing APIs and data requirements among the core PLUTO system components, as well as development of core components (search engine, TM and MT) progressing around shared data and languages for the 12-month prototype system (due April 2011).

Dissemination activities are already well underway with the basic infrastructure of project logo, web site, fact sheet, brochure, presentation etc. being complemented by partner dissemination activities at relevant international conferences and events.

In addition to planned project milestones and deliverables, a major highlight of our achievements during the first calendar year relates to the PLUTO project's engagement on a pilot Machine Translation project for English—Portuguese with the European Patent Office. We were able to successfully develop and deploy a machine translation system with EPO data, pass stringent engineering testing and quality review by the national patent office, and achieve the EPO project officer's recommendation that the system be accepted for use at the EPO's Espacenet service. The system has been in live production use at Espacenet since the beginning of October 2010.

2. Project Overview

A core aspect of the European Commission's (EC) commitment to language diversification is the provision of multilingual access to intellectual property information, namely patents. This will afford inventors in Europe better access to technical information on patents in their native language and foster innovation and growth. Central to such a provision is the availability of high-quality search and translation technologies capable of dealing with the volume and language diversity of large collections of patent data. MT software must also be adapted to handle the specific language found in patent documents.

In order to adequately facilitate this, the PLUTO project aims to develop an online framework – through the integration of a number of mature software components – whereby users can exploit state-of-the-art tools to search large patent repositories and automatically translate the retrieved documents.

As well as directly addressing the translation needs of the EC, PLUTO serves a more general purpose when it comes to IP related activities. SMEs and individual inventors can encounter difficulties when entering a new market due to the high costs related to prior art search and translation. Often, making such a leap constitutes a large risk for these entities. PLUTO aims to reduce the risk by providing an integrated, online translation tool where clients – in the form of several human experts (technical, legal, consultants) – can take advantage of existing web-content and state-of-the-art MT technologies and information retrieval (IR) tools to collaboratively search for, retrieve and translate patents in a fast, cost-effective manner.

In doing this, PLUTO will also advance the state-of-the-art in MT through novel approaches to integration with TM and domain adaptation techniques aimed at dealing with the specific characteristics of patent documents (legalese, technical terminology and long sentences). Furthermore, a number of innovative approaches to cross-lingual information retrieval will allow for more efficient patent searching.

3. Progress to Date

Achievements

Despite the project being in its relative infancy, significant progress has been made towards meeting a number of project milestones and goals. Some key achievements to date include:

- Deployment of an online production-level MT service for English—Portuguese at the European Patent Office;
- Development of a first prototype for the PLUTO web service and patent search engine;
- Specification of an integration framework for the individual software components involved in the overall service.

Other Project-related Work

In addition to the aforementioned achievements and outputs, substantial work was carried out in line with the predefined work plan of the project.

- Patent data was made available to the consortium in a number of languages from two sources: the IRF's MAREC corpus and the European Patent Office (EPO).
- Project partners have defined the XML schema via which documents will be exchanged between the search engine, the translation memory, and the machine translation (MT) system.
- Translation memory resources have been generated from the data supplied from the IRF and the EPO and added to the repository at ESTeam.
- As well as the MT engines for EN—PT, further engines for EN—FR have been developed including investigations into a number of approaches for domain adaptation according to the International Patent Classification system.

Advisory Board

The consortium has confirmed the following individuals as members of our advisory board: Dr. Fred Hollowood, Dr. Gregory Grefenstette, and Viggo Hansen.

4. Dissemination Activies

A dissemination plan for the project was developed as a deliverable over the course of the first six months of the project. Our strategy is to promote the project and its output through a number of channels, including the project website, social media outlets, at conferences through publications, presentations and demos/showcases, and via press releases and other communication materials

The PLUTO website (http://www.pluto-patenttranslation.eu) went live on July 16th 2010. The project also has a Facebook page and a presence on Twitter via the hashtag #PLUTO_EU.

EPO Collaboration

We have been able to capitalise on the collaboration with the EPO as a means to further promote the project and to satisfy some dissemination requirements. Firstly, the technical aspects of the work form the core of the publication accepted at the AMTA conference (see Publications below). Additionally, the service which the consortium provides to the EPO has been publicised on both the website of the project as well as the website of the EPO (http://www.espacenet.com/) and can be referred to in future dissemination materials. We also intend to source testimonials from the EPO and users of the translation engine to promote the project from a commercial stand point.

Publications

We have had one paper accepted to the commercial user track of the AMTA conference (American Machine Translation Association). The paper is publicly available for download from the project website and its reference is given below:

- **John Tinsley, Andy Way and Páraic Sheridan. 2010.** *PLuTO: MT for Online Patent Translation*. In Proceedings of AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas, Denver, CO.

Conferences and Presentations

- **EAMT 2010** (May 26/27th, St-Raphael, France)
- **IRF Symposium 2010** (June 1—4th, Vienna, Austria)
- **SIGIR 2010** (July 19—23rd, Geneva, Switzerland)
- **AMTA 2010** (November 1—4th, Denver, USA)