PLuTO: MT for Online Patent Translation

John Tinsley
CNGL
School of Computing
Dublin City University, Ireland
jtinsley@computing.dcu.ie

Andy Way
CNGL
School of Computing
Dublin City University, Ireland
away@computing.dcu.ie

Páraic Sheridan
CNGL
School of Computing
Dublin City University, Ireland
psheridan@computing.dcu.ie

Abstract

PLuTO – Patent Language Translation Online – is a partially EU-funded commercialization project which specializes in the automatic retrieval and translation of patent documents. At the core of the PLuTO framework is a machine translation (MT) engine through which web-based translation services are offered. The fully integrated PLuTO architecture includes a translation engine coupling MT with translation memories (TM), and a patent search and retrieval engine. In this paper, we first describe the motivating factors behind the provision of such a service. Following this, we give an overview of the PLuTO framework as a whole, with particular emphasis on the MT components, and provide a real world use case scenario in which PLuTO MT services are exploited.

1 Introduction

The European Commission has supported human language technologies, in particular MT, for over 40 years. This has led to a number of pioneering developments in these areas. This support has been particularly concerted in the past decade due to changes in the commercial landscape in Europe, where research indicates that consumers feel constrained to buying only in their own language due to issues with language barriers.

A core aspect of the EU's 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 MT technology adapted to handling the specific language found in patent documents. This technology is increasingly important as the Commission continues their efforts to secure a move towards a single EU patent. The EU patent – filed in one of English, French or German – would introduce translation needs for countries with other official languages. For instance, while an EU patent filed in Italy could be filed in Italian, it would also be required to be translated into either English, French or German.

As part of the Commission's proposal for an EU patent, accompanying measures have been set out in order to make the existing (multilingual) European patent system more accessible to inventors. Again, this requires MT technology capable of dealing with the volume and language diversity of such a collection of data. To adequately facilitate this, the EU has partially funded the PLuTO (Patent Language Translations Online)[1] project to develop a framework whereby users can exploit MT to translate large patent repositories.

In addition to directly addressing the translation needs of the Commission, PLuTO serves a more general purpose when it comes to IP related activities. Small and Medium Enterprises (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 ex-

[1] http://www.pluto-patenttranslation.eu

perts (technical, legal, consultants) – can take advantage of existing web-content and state-of-the-art MT technologies and information retrieval tools to collaboratively search for, retrieve and translate patents in a fast, cost-effective manner.

2 PLuTO Overview

PLuTO is funded for 3 years (beginning April 2010) by the EU under the Information and Communication Technologies Policy Support Programme (ICT-PSP). The main objective of the ICT-PSP is to stimulate innovation in, and commercialisation of, online multilingual services. To this end, projects are only supported for 50% of their costs, meaning that applicants must believe they have a commercially viable solution that they are willing to put significant capital behind before they seek EU support.

Despite the fact that the PLuTO project is still in its relative infancy, we are already involved in large-scale commercial development and deployment activities described further in section 4. However, we must stress at this juncture that the framework discussed in section 6 is not yet a fully commercial system, nor are what we describe in sections 3 and 4 the limits of our MT system's capabilities. There will be significantly more developments and innovations over the lifetime of PLuTO.

2.1 The PLuTO Consortium

PLuTO comprises a dynamic industry-academia consortium, each member of which brings significant experience and expertise to some facet of the service.

The Centre for Next Generation Localisation (CNGL)[2] at Dublin City University brings to the target platform a state-of-the-art MT engine, MaTrEx (Stroppa and Way, 2006; Tinsley et al., 2008, Penkale et al., 2010) (described further in section 3). ESTeam AB contributes a comprehensive translation software environment, including multi-layered, multi-domain translation memory technology. The Information Retrieval Facility (IRF) provides search and retrieval expertise as well as a substantial multilingual patent repository. Finally, CrossLanguage and The Dutch Patent User Information Group (WON) provide evaluation, analysis and feedback on the quality of PLuTO's patent retrieval and translation services from a linguistic and patent user perspective respectively.

2.2 PLuTO Objectives

Given the translation needs of the European Commission and the obstacles to innovation touched on previously, we have identified a number of key objectives for PLuTO:

a) Development of a rapid solution for patent search and translation by drawing on the expertise of our consortium and integrating our existing software components and adapting them to the relevant domains and languages.

b) Commoditisation of domain-adaptable MT, allowing for the high volume translation of existing patent repositories as well as making MT more accessible to SMEs and individual inventors who do not have the resources to contract patent lawyers, examiners and translators. In doing this, we will be simultaneously making MT technologies more visible to potential users.

c) Building a fully-functional indexing engine that will serve as a complete tool for prior art search by professional patent searchers or individual inventors.

d) Deployment of a professional Web 2.0 application where patent searchers, lawyers, translators and others involved in the prior art search process, can collaborate on a particular task, sharing search results and translations in order to achieve and optimal solution.

We expect that PLuTO will ultimately give rise to a more inclusive innovation society where individual inventors and SMEs have access to a large pool of multilingual information without language barriers.

3 Machine Translation in PLuTO

MT in PLuTO is carried out using the MaTrEx [3] (**Ma**chine **Tr**anslation Using **Ex**amples) system developed at DCU. It is a hybrid data-driven systems built following established design patterns,

[2] http://www.cngl.ie

[3] http://www.openmatrex.org

with an extensible framework allowing for the interchange of novel or previously developed modules. This flexibility is particularly advantageous when adapting to new language pairs and exploring new processing techniques, as language-specific components can be plugged in at various stages in the translation pipeline.

The hybrid architecture has the capacity to combine statistical phrase-based, example-based and hierarchical approaches to translation. MaTrEx also acts as a wrapper around existing state-of-the-art components such as Moses (Koehn et al., 2007) and Giza++ (Och and Ney, 2003). Subsequent novel development of the system has resulted in the MaTrEx system achieving world leading ranking in English–French, Spanish and Chinese MT at the Workshop on Statistical Machine Translation (WMT-09) as well as for non-EU languages, such as first place for Arabic–English at IWSLT-07 and for English–Hindi at ICON-08. The principal implemented components of the MaTrEx system to date include: word alignment (including word packing), chunking, chunk alignment, treebank-based phrase extraction, supertagging, and decoding. The system also includes language-specific extensions such as taggers, parsers, etc. used in pre- and post-processing modules. All of these modules can be plugged in or out, depending on the needs of the language pair and translation task at hand.

3.1 MT at Production Level

There are both advantages and disadvantages to using data-driven models for producing MT systems on this scale. The most obvious benefit is the potential for rapid deployment given requisite training materials. Large-scale systems can be built from the ground up in a matter of days. Additionally, as training of the engines is performed offline, we can deploy early versions of systems which are then iteratively improved and released as updated versions .

An obvious drawback of data-driven system is the need to procure significant amounts of relevant training material. However, for PLuTO, this is not an issue: in fact, quite the opposite. Given the large amounts of training data at our disposal, through our partners and clients, our main difficulty is dealing with large models – translation, reordering and language – in an efficient manner. Large amounts of training data produce large models which, while producing better translation results, are very resource-intensive and can be quite slow at translation (decoding) time.

Typically, in the research world, this is overcome by filtering the models based on the input to be translated. However, as we are dealing with a real-world scenario, we do not have the luxury of knowing what the input is going to be. Thus, finding a satisfactory compromise between the three key ingredients – quality, speed and computational resources – is core to the success of PLuTO. We discuss some ways in which we tackle this issue in section 4.2.

3.2 Deployment of MT Services

Accessing PLuTO MT services can be envisaged in a number of ways. The system can be hosted on PLuTO servers and accessed remotely or, potentially, the services could be customised to allow them to be hosted locally on a client's server.

In the fully integrated PLuTO system (described further in section 6), translation is carried out as a backend to search functionality, i.e. users will search for a patent and, once retrieved, have the option of translating it via a "translate" button in the web interface.

Alternatively, translation can be provided directly through an API, similar to Google Translate, where the user can specify the document type, patent code and language pair.

3.3 Data Processing

From the time a document is submitted for translation, aside from actual MT, there is significant data processing involved in the end-to-end pipeline. Documents can be input in a number of different formats – plain text, XML, HTML, and pdf – all of which must be checked and verified for correctness before proceeding. Following this, the processing steps required to prepare the data in the format expected by the MT system must be carried out. Following translation, the data must be converted back into the original format supplied by the user. Each of these steps presents its own difficulties, yet must be carried out quickly and efficiently so as to not introduce any significant lag in translation time as visible to the user.

3.4 Patent Translation

Patent translation is a unique task given the nature of the language used to produce them. Documents contain a mixture of legal vernacular and scientific and specific terminology related to the topic of the patent. Because of this, the task of building an MT system for patents is not as straightforward as collecting masses of parallel data and training a system. Rather, we have to carefully select relevant in-domain[4] data from our patent collections and potentially adapt multiple MT systems for different sets of patents. Fortunately, there exists a hierarchical patent classification system – the International Patent Classification (IPC) – created by the World Intellectual Property Organisation (WIPO), which greatly assists us in this task

3.4.1 IPC Codes

According to the WIPO, the IPC "provides for a hierarchical system of language-independent symbols for the classification of patents and utility models according to the different areas of technology to which they pertain".[5] On the top level of the IPC, patents are classified according to one of 8 categories and assigned a letter, including:

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

The hierarchy then becomes very granular and patents are assigned additional symbols relating to the technology they describe. As patent documents are filed with their relevant IPC codes, all patent data we have for training MT systems is preclassified. We can then chose the most appropriate level of classification specificity when building our engines.

Additionally, having this information at hand allows us to access IPC code-specific dictionaries we may have, and/or, depending on the training procedure employed when building the MT engine, allow us to access domain-specific translation and language models which may provide us with more appropriate translations given the input.

3.4.2 Domain Adaptation

From our limited experience, it is still an open question as to the best way to exploit in-domain data for MT. Using just the top level of the IPC, we have the capability to build 8 in-domain translation and language models. However, depending on the distribution of our patent data, we may have varying quantities of training materials for each domain. For example, when dealing with one particular client, they provided us with their own patent-based translation memories with which to train our MT engines. For domain C of the IPC, we had ~2.2M training pairs, while for domain E we had only ~59,000 pairs. This begs the question whether we should really build individual in-domain systems for each code, and risk data spareness in those under-resourced ones. We could potentially build more general systems by combining patent data from similar domains (based on some similarity measures) considering, for example domains G and H, to be close enough to combine. Another option is to just combine all data and build a single 'general' (patent) domain system.

When building a general system, one might then reasonably ask the question as to how useful phrase pairs from domain C are when translating data from domain E. Intuitively, we would think not very helpful, yet based on our experimentation to date, this has not necessarily been the case. For illustrative purposes, in the following section we describe a portion of a set of experiments we have carried out related to investigating the optimal configuration of domain-adapted MT for patents.

3.5 Domain Adaptation - Experiments

In this section, we describe some translation experiments we carried out from Portuguese into English using our MaTrEx MT engine. The idea behind these experiments was to give us a sense as to the best way to build MT systems for patents given data from multiple domains. We selected three patent domains based on the IPC codes – C, B, E – given that these domains were the ones we

[4] By in-domain here, we do not mean in the patent domain, but rather the sub-domains within the patent domains.
[5] http://www.wipo.int/classifications/ipc/en/

had 1) the most training data for, 2) the least training data for, and 3) somewhere in between. Additionally, while the type of language found in domains B and E can often resemble regular natural language with specific terminology, domain C has large amounts of chemistry specific chunks, including chemical names and complex formulae that one could not envisage being used in any other context. The exact details for each domain are given in Table 1.

Domain	# Training Pairs
C	2,214,077
B	441,030
E	59,050

Table 1. Quantities of training data for different domains

For each domain, we had a testset comprising 1,000 sentences and we built a number of systems combining in-domain data with data from the other domains. For instance, for Testset B, if we have t-table (translation model) as “in-domain” and l-model (language model) as “general”, that means we built a translation model using only data from domain B and a language model combning data from all three domains; B, C and E. Thus, for each testset we built systems using all four possible combinations of in-domain and general translation and language models. Translation quality was measured using the BLEU and METEOR metrics. The results of these experiments for each of the three domains are given in Tables 2—4.

Testset B (Performing Operations...)			
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
General	General	0.5495	76.53

Table 2. Results from Testset B

Testset C (Chemistry...)			
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
General	General	0.5998	80.74

Table 3. Results from Testset C

Testset E (Fixed Constructions)			
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
General	General	0.5679	77.44

Table 4. Results from Testset E

Our findings are quite inconclusive when it comes to determining how best to exploit in-domain data. Looking at Table 2, we see that simply adding general data to the language model gives significant improvements, while adding even more to the translation model improves things further. We expect that, despite the fact that the data in domain C can be quite chemistry-specific, given its sheer volume, there is sufficient natural language common to that found in domain B to give rise to improved translations.

Looking at Table 3, we see that adding general domain data to the language model does not help. The purpose of the language model in MT is to improve fluency in the target-language output. Given that the output from domain C would often be very specific chemistry-related language, the more general data from the other domains was not of much use in helping to make this output more fluid. However, we see that adding the general data to the translation model gives significant improvements. The general data is useful in this case as it helps in terms of informing the translation selection process, but again we see using a general language model on top of this gives no further improvements.

Finally, looking at Table 4, we see that any configuration which employs general data sees an improvement. This is due to the relatively small size of our training set for domain E. Consequently, when we add additional data regardless of the domain, translation quality improves.

Overall, these experiments indicate that there are benefits to be had in some respects by exploiting data from other patent domains. However, which domains best combine, where in the translation system (translation or language model) this data is most effective, and how much is needed for it to be effective, are questions which require further investigation. It is one of our goals over the

lifetime of the PLuTO project to ascertain the optimal configurations of domain-adapted MT systems in this regard.

4 Use Case: Online MT

PLuTO MT is currently used as a backend to a high-volume patent search and translation facility. Translation services are provided 24/7 through a web service whereby users search for patent documents and select the sections – abstract, claims, description – they want to translate. A secure connection is established between the client and server to ensure that the translation services are not exploited by unauthorised users.

4.1 Data Processing and Load Management

When users make a request for translation, the document to be translated is sent by calling our REST API with a URL. This is the process to determine the IPC code of the document, the language pair and other information so that the most appropriate translation engine can be identified. The text to be translated is then extracted from the document, preprocessed and distributed to the translation engines across our servers.

In order to return translations as quickly as possible, we use multiple multi-core processors to translate multiple sentence simultenously. We carry out careful load balancing based on the length of the input sentence and the expected time it will take to translate, in order to allow our task farming proceedures to distribute the translations in such a way that ensures optimal performance.

At present, the demand placed on our servers is approximately 10,000 translation requests (of full XML documents) per month which we translate at an estimated average rate of 27 words per second[6] on a single processor. Exploiting our taskfarming capabilities fully, we can increase our translation rate to approximately 280 words per second.

4.2 MT Model Management

For this particular task, PLuTO was provided with a large amount of data for the purposes of training our MT system including TMX files, OLIF dictionary files and monolingual pdf and XML documents. A summary of quantities of data is given in Table 5.

Data	Quantity
TMX segments	5.5M (approx)
Dictionary entries	721,360 (all domains)
Monolingual data (words)	102,085,965

Table 5. Summary of training data for online MT system

As we mentioned in section 3.1, while having such abundant training resources is desirable in terms of the quality of translations we are able to produce, it raises further questions: given such training data, how do we produce models of a manageable size that return translations in a reasonable amount of time?

Simply training a system using our standard configuration resulted in models of unmanageable size: >250M phrase pairs in the translation model, and >260M *n*-grams in the language model. Even if we had enormous quantities of RAM in which to fit these models or carried out some tricks to load them on demand, translation would still be too slow. In order to establish optimal MT engine configuration, we carried out extensive testing prior to deploying our system. We built a number of different systems by pruning the phrase table to different degrees, building language models of different orders and employing different methods of loading the models into memory. As mentioned in section 3.1, there were three key factors to consider when evaluating these systems: translation quality, speed of translation and memory requirements. We measured these and plotted them against each other to select the optimal system for this particular task. Our results are illustrated in Figure 1. In this figure, each bubble represents a particular MT system configuration and the larger the bubble, the larger its memory footprint (the memory values are given beside each bubble). Ideally, in this graph, we would have a small bubble in the upper right corner. We can see from Figure 1 that choosing the optimal system is not straightforward. It will always be dependent on the particular task, namely whether more priority is given to translation speed or quality, and also the resources at ones disposal.

[6] Estimates are based on an average translation time across all the language pairs for which we provide translations.

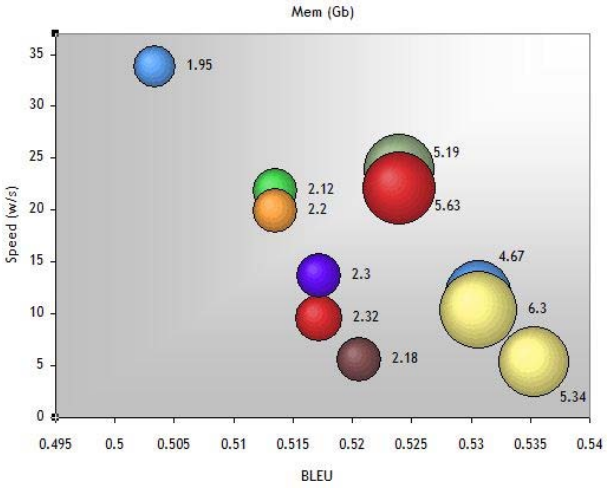

Figure 1. Balance between quality, speed and memory consumption. The size of the bubbles relates to the MT systems memory footprint (values are also provided beside each one)..

5 Patent Search and Retrieval in PLuTO

PLuTO search is performed using cross-lingual information retrieval based on the Solr/Lucene open-source platform. This open-source platform already supports multilingual search, providing different analysers and stemmers for a variety of languages. Our large scale patent repository is indexed and many of the documents already occur in more than one language. This allows us to bypass the MT stage when a user requests a translation of a document, and simply retrieve the section in the relevant language.

For example, if we have patent document US2010082324 in both English and Spanish, we will have a single XML document with multiple instances of the same element for both languages as illustrated below:

```
<description lang=EN>
Machine translation systems…
</description>

<description lang=ES>
Sistemas de traducción
automática…
</description>
```

When requests to translate such documents are made, it is first checked whether an element exists for the chosen target language. If it does, then it can be returned as a translation immediately, otherwise it is passed to the MT system.

6 Integrated System Architecture

In the following, we describe some of the (planned) components and functionality of the fully integrated PLuTO service on a relatively high-level for illustrative purposes as to our intentions going forward.

6.1 MT and TM Functional Integration

Internally the TM and MT modules constitute self-standing entities and it is the function of the integration module is to select the process which offers better quality, scalability and portability to new domains. There are several ways in which this functional integration can be configured based on the nature of the translation task.

The various scenarios can operate on the sentential and/or sub-sentential level. One option is to present the segments to the TM module and if a (fuzzy) match of predefined quality is not found, the segments are passed to the MT system. We also have the option of employing more sophisticated means of integration, e.g. He et al. (2010).

6.2 Search and Translation Integration

The integration of search and translation is just a simple interaction whereby users can chose to send certain elements of patent search results, e.g. entire patent document, abstract, claims, etc., to the MT engine for translation.

6.3 PLuTO Web Application

The web application is the hub through which all users of the integrated PLuTO service interact with the search and translation tools. This solution concentrates on simplicity and user friendliness, where users are able to submit documents for translation or search the existing multilingual repository, select, translate and collaborate towards a patent search report. The web application as a whole comprises a number of distinct modules:

1. **Web Utility/Interface:** This is the main interface of PLuTO which serves as a collaborative environment where multiple users can login simultaneously to work on

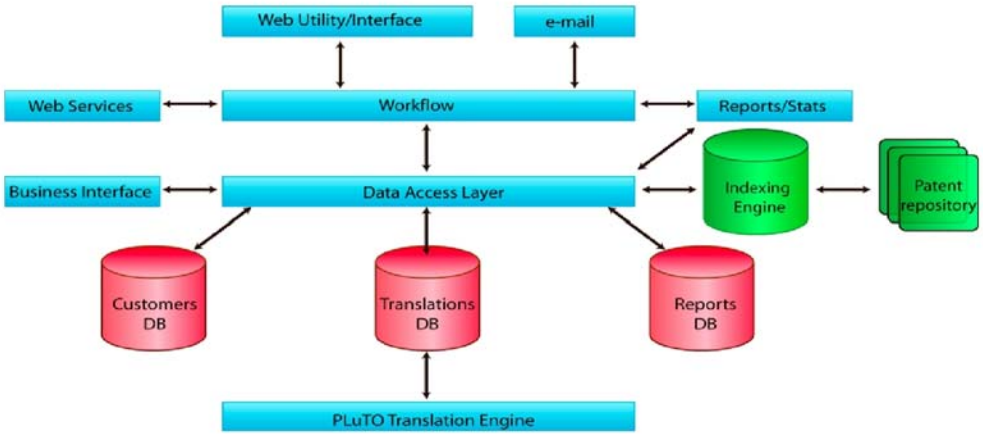

Figure 2. Overview of full (proposed) PLuTO architecture

a single project and access the search and translation facilities.

2. **E-mail Management:** This module allows users to share patent documents and translations directly from the PLuTO interface via email.
3. **Reports/Statistics:** This is a standalone library through which users can view statistics related to their project, e.g. number of documents retrieved etc. It also serves as a space in which patent reports can be collaboratively worked on.

An overview of the full PLuTO architecture is shown in Figure. 2.

7 Conclusions

We have described the PLuTO project from its origins and discussed various technological aspects involved, with particular focus on MT. In the context of the use case in which our translation technology is currently being exploited, we have shown how one much strike a delicate balance between translation quality, speed of translation and required computational resources, particularly when abundant training data is available.

We have also demonstrated that the issue of domain adaptation for patent translation in non-trivial. It may be the case that optimal configurations must be ascertained on a case by case basis given the training data available for the task.

Finally, we have given an overview of the full proposed architecture for the project and outlined the plans for PLuTO over the coming years.

Acknowledgments

The PLuTO Project (ICT-PSP-250430) is generously supported under the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme.

References

He, Y., Yanjun Ma, Andy Way and Josef Van Genabith. 2010. Integrating N-best SMT Outputs into a TM System. In *Proceedings of COLING 2010*, Beijing,

Penkale, S., Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, Andy Way. 2010. MaTrEx: The DCU MT System for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, ACL 2010*, Uppsala, Sweden, pp.143–148.

Stroppa, N. and A. Way. 2006. MaTrEx: DCU Machine Translation System for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, pp.31–36.

Tinsley, J., Y. Ma, S. Ozdowska and A. Way. 2008. MaTrEx: the DCU MT System for WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation, ACL 2008*, Columbus, OH.

China (forthcoming).