PLUTO: Automated Solutions for Patent Translation[i]

John Tinsley, Alexandru Ceausu, Jian Zhang

Centre for Next Generation Localisation
School of Computing
Dublin City University, Ireland
{jtinsley;aceausu;jzhang}@computing.dcu.ie

1 Introduction

PLUTO is a commercial development project supported by the European Commission as part of the FP7 programme which aims to eliminate the language barriers that exist worldwide in the provision of multilingual access to patent information. The project consortium comprises four partners: the Centre for Next Generation Localisation at Dublin City University,[1] ESTeam AB,[2] CrossLang,[3] and the Dutch Patent Information User Group (WON).[4] Research and development is carried out in close collaboration with user groups and intellectual property (IP) professionals to ensure solutions and software are delivered that meet actual user needs.

1.1 The need for patent translation

The number of patent applications filed worldwide is continually increasing, with over 1.8 million new filings in 2010 alone. Yet Despite the fact that patents are filed in dozens of different languages, language barriers are no excuse in the case of infringement. When carrying our prior-art and other searches IP professionals must ensure they include collections which encompass all potential relevant patents. Such searches will typically return results – a set of patent documents – 30% of which will be in a foreign language.

As professional translation for patents is such a specialist task, translators command a premium fee for this service, often up to €0.50 per word for Asian languages. This often results in high or unworkable translation costs for innovators. While free machine translation (MT) tools such as Google translate have unquestionably been beneficial in helping to reduce the need to resort to expensive human translation, the quality is still often inadequate as the models are too general to cope with the intricacies of patent text.

In what follows, we will provide an overview of some of the technologies being developed in PLUTO to address the need for higher quality MT solutions for patents and how these are deployed for the benefit of IP professionals.

2 Language Technology for Patents

Patent translation is a unique task given the style of language used in patent documents. This language, so-called "patentese", typically comprises a mixture of highly-specific technical terminology and legal jargon and is often written with the express purpose of obfuscating the intended meaning. For example, in 2001 an innovation was granted in Australia for a "Circular Transportation Facilitation Device", i.e. a wheel.[5]

Patents are also characterised by a proliferation of extremely long sentences, complex chemical formula, and other constructs which make the task for MT more difficult.

2.1 Domain-specfic machine translation

The patent translation systems used in PLUTO have been built using the MaTrEx MT framework (Armstrong et al., 2006). The systems are domain specific in that they have been trained exclusively using parallel patent corpora. A number of experiments related to domain adaptation of the language and translation models have been carried out in the context of these systems. The principal findings from this work were that systems combining all available patent data for a given language were preferable (Ceausu et al. 2011).

[1] www.cngl.ie
[2] www.esteam.se
[3] www.crosslang.com
[4] www.won-nl.com

[5] http://pericles.ipaustralia.gov.au/aub/pdf/nps/2002/0808/2001100012A4/2001100012.pdf

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 69–71,
Avignon, France, April 23 - 27 2012.

Significant pre-processing techniques are also applied to the input text to account for specific features of patent language. For instance, sentence splitting based on the marker hypothesis (Green, 1979) is used to reduce long sentences to more manageable lengths, while named-entity recognition is applied to isolate certain structures, such as chemical compounds and references to figures, in order to treat them in a specific manner.

Additionally, various language-specific techniques are used for relevant MT systems. For example, a technique called word packing (Ma et al., 2007), is exploited for Chinese—English. This is a bilingually motivated task which improves the precision of word alignment by "packing" several consecutive words together which correspond to a single word in the corresponding language.

Japanese—English is a particularly challenging pair due to the divergent word ordering between the two languages. To overcome this, we employ preordering of the input text (Talbot et al. 2011) in order to harmonise the word ordering between the two languages and reduce the likelihood of ordering errors. This is done using a rule-based technique called head-finalisation (Isozaki et al., 2010) which moves the English syntactic head towards the end of the phrase to emulate the Japanese word order.

Finally, we use compound splitting and true casing modules for our English—German MT systems in order to reduce the occurrence of out-of-vocabulary words.

2.2 Translation memory integration

In order to further improve the translation quality, we are developing an engine to automatically combine the outputs of the MT system and a translation memory (TM).

The engine works by taking a patent document as input and searching for full matches on paragraph, sentence, and segment (sub-sentential) level in the TM. If no full matches are found, fuzzy matches are sought above a predetermined threshold and combined with the output of the MT system using phrase- and word-level alignment information.

For patents, most leverage from the TM is seen at segment level, particularly as the patent claims are often written using quite a rigid structure. This is due to that fact that, as patents typically describe something novel which may never have been written about previously, there is often little repetition of full sentences.

2.3 Evaluation

The performance of the patent MT systems in PLUTO is evaluated using a range of methods aimed not only at gauging general quality, but also identifying areas for improvement and relative performance against similar systems.

In addition to assessing the MT systems using automatic evaluation metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee et al. 2005), large-scale human evaluations are also carried out. MT system output is ranked from 1—5 based on the overall quality of translation, and individual translation errors are identified and classified in an error categorisation task.

On top of this standalone evaluation, the PLUTO MT systems are also benchmarked against leading commercial systems across two MT paradigms: Google Translate for statistical MT and Systran (Enterprise) for rule-based MT. A comparative analysis is carried out using both the automatic and human evaluation techniques described above. This comparison is also applied to the output of the PLUTO MT systems and the output of the integrated TM/MT system in order to quantify the improvements achieved using the translation memories.

The main findings from the first round of evaluations for our French—English and Portuguese—English systems showed that our MT systems score relatively high based on human judgments -- 3.8 out of 5 on average -- while being ranked higher than the commercial systems approximately 75% of the time. More details on these experiments can be found in Ceausu et al. (2011).

3 Patent Translation Web Service

The PLUTO MT systems are deployed as a web service (Tinsley et al., 2010). The main entry point for end users is through a web browser plugin which allows them to access translations on-the-fly regardless of the search engine being used to find relevant patents. In addition to the browser plugin, users also have the option to input text directly or upload patent documents in a number of formats including PDF and MS Word.

A number of further natural language processing techniques are exploited to improve the user experience. *N*-gram based language identification is used to send input to the correct MT system; while frequency based keyword extrac-

tion provides users with potentially important terms with which to carry out subsequent searches.

Corresponding source and target segments are highlighted on both word and phrase level, while users have the option of post-editing translations which are stored in a personal terminology database and applied to future translations.

The entire framework has been designed to facilitate the patent professional in their daily workflow. It provides them with a consistency of translation quality and features regardless of the search tools being used to locate relevant patents.

This has been validated through extensive user experience testing which included a usability evaluation of the translation output.

4 Looking Forward

The PLUTO project has been running for just over two years and is scheduled to end in March 2013. Our goal by that time is to have established a viable commercial offering to capitalize on the state-of-the-art research and development into automated patent translation.

In the meantime, we will continue to build upon our existing work by building MT systems for additional language pairs and iteratively improving upon our baseline translation performance. Significant effort will also be spent on optimising the integration of translation memories with MT using techniques such as those described in He et al. (2011).

Acknowledgements

The PLuTO project (ICT-PSP-250416) is funded under the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme

References

Armstrong, S., M. Flanagan, Y. Graham, D. Groves, B. Mellebeek, S. Morrissey, N. Stroppa and A. Way. 2006. *MaTrEx: Machine Translation Using Examples.* TC-STAR OpenLab on Speech Translation. Trento, Italy.

Banerjee, S. and Lavie, A. (2005). *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-05), Ann Arbor, MI.

Ceausu, Alexandru, John Tinsley, Andrew Way, Jian Zhang, Paraic Sheridan, *Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project*, The 15th Annual Conference of the European Association for Machine Translation, EAMT-2011, Leuven, Belgium

Green, T., *The necessity of syntax markers. two experiments with artificial languages*. Journal of Verbal Learning and Behavior, 18:481{496}, 1979.

Isozaki, H., Sudoh, K., Tsukada, H., and Duh, K. *Head finalization: A simple reordering rule for SOV languages.* In Proceedings of the 5th Workshop on Machine Translation (WMT), Upsala, Sweden.

Ma, Yanjun, Nicolas Stroppa, and Andy Way. 2007. Boostrapping Word Alignment via Word Packing. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, pp.304—311

Papineni, K., Roukos, S.,Ward, T., and Zhu,W.-J. (2002). *BLEU: a Method for Automatic Evaluationof Machine Translation*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), pages 311–318, Philadelphia, PA.

Talbot, David, Hideto Kazawa, Hiroshi Ichikwa, Jason Katz-Brown, Masakazu Seno, Franz Och *A Lightweight Evaluation Framework for Machine Translation Reordering*, In Proceedings of the Sixth Workshop on Statistical Machine Translation (July 2011), Edinburgh, Scotland.pp. 12-21

Tinsley, J., A. Way and P. Sheridan 2010. *PLuTO: MT for Online Patent Translation* In Proceedings of the 9th Conferences of the Association for Machine Translation in the Americas. Denver, CO, USA.

[i] This paper is an extended abstract intended to accompany an oral presentation. It is not intended to be a standalone scientific article.