Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project

Alexandru Ceauşu, John Tinsley, Jian Zhang, Andy Way

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

Abstract

The PLUTO[1] 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 project partners. The paper presents some of the experiments on patent domain adaptation of the Machine Translation (MT) systems used in the PLuTO project. The experiments use the International Patent Classification for domain adaptation and are focused on the English–French language pair.

1 Introduction

The European Commission has supported human language technologies, in particular Machine Translation (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 Commission'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 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. To this end, the European Commission has part-funded the PLuTO (Patent Language Translations Online) project to develop a framework in which their users can exploit state-of-the-art MT to translate patent documents.

As well as supporting the translation needs of the Commission, PLuTO serves a more general purpose when it comes to intellectual property-related activities. There are considerable translation requirements throughout the end-to-end patent application process. The necessary quality and quantity of translations varies greatly depending on the stage in the process. For example, at the patentability/prior-art searching stage, dozens of documents need to be translated but the quality does not need to be perfect; on the contrary, when establishing freedom to operate, a small number of documents must be precisely translated as there are legal implications involved.

At present, there are a limited number of tools that can carry out such translations adequately; at least not for what might be deemed an economical price. Small- and Medium-size Enterprises and individual inventors can encounter difficulties when entering a new market due to the high costs related to translation. Often, making such a leap constitutes a large risk for these entities. Additionally, local patent agencies – who typically provide expert patent translation services – are overburdened with requests for human translations.

The PLuTO project aims to support these different users by developing a number of tools –

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

including an online framework which integrates a number of mature software components – with which users can facilitate their patent search and translation needs.

In doing this, PLuTO will also advance the state-of-the-art in MT through novel approaches to integration with translation memory (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 techniques will be developed to allow users to incorporate MT into their patent search workflows.

In this paper, we present some experiments carried out to date on patent domain adaptation for MT. Domain adaptation offers two opportunities for MT improvement: (i) it might be regarded as the task of adapting the MT system to the particular style of language used in patent documents, and (ii) if separate MT systems are used for each patent area of technology, then the general MT system accuracy might improve, as shown in (Banerjee et al., 2010).

The remainder of the paper in organised as follows: the second section gives an overview of the PLuTO MT system technology and architecture, as well as providing details on the data preparation stage for patent translation. In section 3 we present the experiments on patent domain adaptation for the English–French translation pair, while in section 4 we present a comparative analysis of the PLuTO system against two commercial systems. Finally, we conclude in section 5.

2 Machine translation in PLuTO

MT in PLuTO is carried out using the MaTrEx (**Ma**chine **Tr**anslation Using **Ex**amples) system developed at DCU (Stroppa and Way 2006; Stroppa et al., 2006; Dandapat et al., 2010). It is a hybrid data-driven system built following established design patterns, 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, 2002). Subsequent novel development of the system has resulted in the MaTrEx system achieving world leading ranking in diverse machine translation shared tasks for language pairs as English–Spanish, English–French (Penkale et al., 2010; Tinsley et al., 2008), as well as for non-EU languages (Almaghout et al., 2010; Okita et al., 2010; Srivastava et al., 2008).

The principal implemented components of the MaTrEx system to date include: word alignment through word packing (Ma et al., 2007), marker-based chunking and chunk alignment (Gough and Way, 2004), treebank-based phrase extraction (Tinsley and Way, 2009), super-tagging (Hassan et al., 2007), 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.

2.1 System architecture

The PLuTO MT framework is currently implemented as a fully-functional web service whereby users can request translations via a number of means, e.g. direct text-based translation through a GUI; as backend to a search result; or by means of a number of bespoke tools. A secure connection is established between the client and server to ensure that the translation services are not exploited by unauthorised users.

The MT system is deployed at the Centre for Next Generation Localisation in Dublin City University as a multi-tier application encompassing three levels:

1. Main access point for patent document translation;
2. Translation server(s);
3. Worker/Decoder server(s).

Communication to and between each of these levels is carried on using XML-RPC conformant messages.

The main access point for patent document translation offers synchronous communication to the MT server through a URL that contains the translation direction. It takes as input an XML document with a format agreed between project partners. The document has bibliographic information (like document number, IPC domains, country, etc.) and at least one of the patent sec-

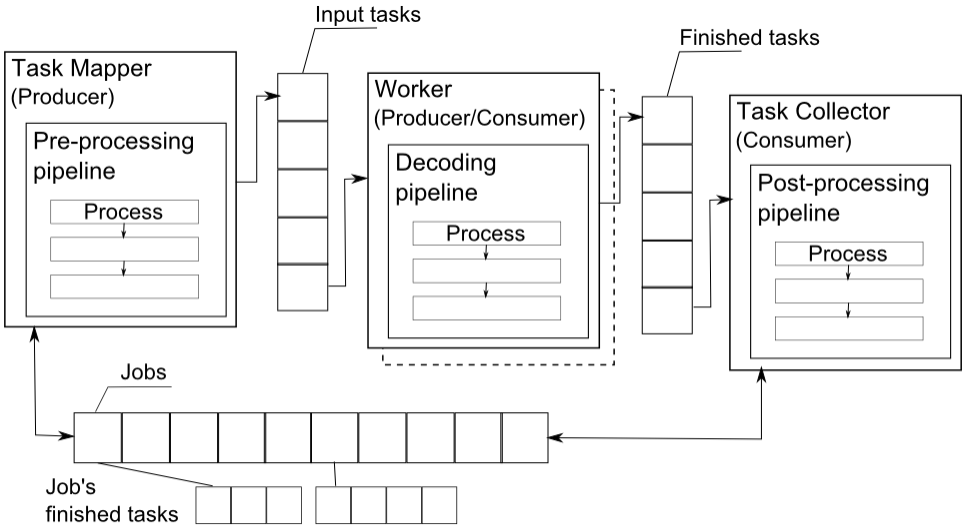

Figure 1. PLuTO MT server multiple producers/consumers architecture

tions (title, abstract, claims or/and description) as shown in example (1).

```
<world-patent-data>
  <fulltext-documents>
    <fulltext-document fulltext-
    format="text-only" system="Pluto
    TM">
     <bibliographic-data>
       <publication-reference data-
       format="docdb">
         <document-id>
           <kind>A1</kind>
         </document-id>
       </publication-reference>
     </bibliographic-data>
     <description lang="en">
       <p>The invention further
       concerns a cosmetic treatment
       method for the skin using said
       composition.</p>
     </description>
   </fulltext-document>
  </fulltext-documents>
</world-patent-data>
```

(1)

The output is the translation of the document in the desired language. The translated document might optionally contain alignment information between source and target at both sentence and token level – example (2).

```
<world-patent-data>
  <fulltext-documents>
    <fulltext-document fulltext-
    format="text-only" system="Pluto
    TM">
     <bibliographic-data>
       <publication-reference data-
       format="docdb">
         <document-id>
           <kind>A1</kind>
         </ document-id>
       </publication-reference>
     </bibliographic-data>
     <description lang="fr">
       <p>L'invention concerne en outre
       un procédé de traitement
       cosmétique de la peau mettant en
       oeuvre ladite composition. </p>
     </description>
     <alignment>
     <sen src="0-94" trg="0-105">
       <seg src="0-12" trg="0-10"/>
       <seg src="14-31" trg="12-31"/>
       <seg src="33-57" trg="33-64"/>
       <seg src="59-81" trg="67-92"/>
       <seg src="83-93" trg="94-104"/>
       <seg src="94-94" trg="105-105"/>
     </sen>
     </alignment>
   </fulltext-document>
  </fulltext-documents>
</world-patent-data>
```

(2)

The main access point for patent document translation transforms each document in a job for the XML-RPC translation servers. Each paragraph from the documents is sent as an asynchronous translation request to the server registered for the given translation direction. There are several XML-RPC methods that provide the asynchronous characteristic of the request:

- `submit_translation` sends a portion of text to be translated (usually a paragraph)
- `request_translation` returns the translation if it is ready or an estimated number of milliseconds to wait for the translation
- `request_alignment` returns the alignment information if the translation is ready or an estimated number of milliseconds to wait for.

In order to return translations as quickly as possible, the translation server has to distribute translation tasks across several cores/machines.

The PLUTO MT system diagram in Figure 1 shows how the system carries on translating multiple sentences simultaneously. The server is based on the multiple producers/consumers pattern. It has a task mapper in which, from a given input text, separate tasks are produced. In our case, the task mapper splits the input into several sentences. There can be one or more workers that pre-process, translate and post-process the translation. The task collector reorders the tasks and delivers the final translation. In-between the task mapper, the workers and the task collector, there are blocking task queues. These queues have prioritization allowing the system to provide a fair-scheduling mechanism for the documents to be translated. That means that each job (document) submitted to the translation server get approximately the same share of the server resources over time. A short document won't have to wait for the completion of a larger document – the sentences from the small document have a higher priority in the workers queue. The workers queue is also capacity-constrained allowing the system to degrade "gracefully". That means that the system won't take more jobs that it can handle in a given time-frame.

All of the server modules are fully configurable through standardized XML files. The same pipelined architecture is shared among workers, task mapper and collector. In this scenario, a pipeline might consist of several processors, with each having serialized initialization and processing functions.

2.2 Data preparation

For the English–French language pairs, the majority of the MT system training data consists of the MAREC-IRF[2] corpus. The MAREC corpus is provided by the Information Retrieval Facility (IRF) and it is the first standardized patent data corpus.

It comprises more than 650GB of multilingual patent documents sourced from the European Patent Office, the World Intellectual Property Organisation, the US Patent and Trademark Office, and the Japan Patent Office. The patent documents of the MAREC corpus have a standardized XML format and they are classified according to the International Patent Classification (IPC).

All patents documents – including those in MAREC – are composed of a *title*, an *abstract*, a *description* (a specification of the patent), a drawing (if it is relevant to the patent) and one or more *claims*. The abstract is the summary of the invention and it is usually around 200 words in length. The description section covers matters such as: the area of the invention; the prior art (previous publicly available information relevant to the originality of the described invention); a sufficient disclosure of the invention; the description of the drawing; and the industrial applicability, amongst other details. Each claim in the claims section is expressed in a single sentence containing three parts: a preamble identifying the domain of the invention (e.g. 'device', 'apparatus', etc.); a transitional phrase that shows how the introductory phrase relates with the content of the claim (e.g., 'comprising', 'consisting', 'including', etc.); and the body of the claim in which the inventor claims a legal monopoly over the invention.

In order to train the MT system for the English–French language pair, we extracted all relevant documents from MAREC. A summary of this data is given in Table 1.

	English	French	Parallel
Abstract	16.57	1.68	1.65
Claims	14.91	7.70	7.56
Description	7.85	0.20	0

Table 1 MAREC English–French document sections used as MT training data (millions)

The majority of the documents with French sections also have an English equivalent. This is not the case with the English documents, where only 10% of the abstracts and 50% of the claims have an equivalent French section, while there are no comparable sections for descriptions across the two languages.

Data preparation for MT training included a number of understated processing steps to clean the data, for example deleting duplicate data, removing lines of text that are in other languages, removing lines or tokens of more than a specified character length, and character encoding normalisation.

In order to create a parallel corpus, the processing stages of sentence splitting and alignment, and tokenisation had to be adapted to the style founding patents. These processes have a number of shared resources such as abbreviations, segmentation rules, and token merging rules. The resources were adapted to the patent language specifics by adding abbreviations that are frequent in patent documents or by adding

[2] http://www.ir-facility.org/prototypes/marec

rules to preserve special types of formulae or chemical compounds.

Following the removal of overly long sentences and pairs with a token ratio of greater than 9:1, we were left with approximately 6 million sentence pairs for training.

3 Domain Adaptation for Patents

Patent translation is a unique task given the nature of the language found in patent documents. Patents typically contain a mixture of legal vernacular and scientific and specific terminology related to the topic in question. Because of this, the task of building MT engines for patents is not as straightforward as collecting masses of parallel data and training a system. In this section, we present some of the techniques we employ when dealing with patents and describe some experiments we carried on domain adaptation using the English–French MAREC corpus.

3.1 Patent-Specific Processing

Aside from the linguistic vagaries of patents, an MT system must also consider the various stylistic and formatting peculiarities. One such characteristic is the propensity to use long sentences which can introduce difficulties for the MT system e.g. long-range reordering. Tokenisation is another non-trivial task in the case of patent documents. Formulae, references to the elements in accompanying figures, references to scientific revues and other patents, and abundant parentheses are just a few of the cases which must be handled with care during tokenisation. In the following, we give two examples of adaptations to the MT engine to handle patent specific characteristics.

References to elements in figures

References to elements in figures are not explicitly difficult to translate: "(1)" typically translates as "(1)". However, there are two less obvious associated problems given the complexities of the MT system: (i) they might be dropped in the translation output because the sequence of words followed by parentheses and numbers has high language model perplexity, and (ii) the individual tokens may get reordered amongst themselves.

Figure references are typically unique to the document in which they occur and thus are unlikely to be observed in the language model. Phrase-based translation can account for local reordering phenomena, but longer word reordering is handled by a separate reordering model. For efficiency, the reordering usually occurs in a limited window of tokens and spurious tokens, such as figure references, often invalidate the longer range reordering mechanism.

In the following example (3), the language model does not account for the trigram "leg (16", and the seventh token in the sequence "(16 , 17 , 18)" – the closing parenthesis – falls outside the default reordering window of six tokens.

```
Preferably , there is more than
one leg ( 16 , 17 , 18 ) that is
attached to the bottom of the          (3)
base member ( 12 ) .
```

The solution we adopted applies a number of rules as a pre-processing step to (a) extract the figure references from the source sentence, (b) translate the sentence without them, and (c) reinsert the references into the correct place based on alignment information stored during decoding.

Long sentences

Long sentences are abundant in patent documents. The most problematic area is the claims section in which the inventor must claim in a single sentence a legal monopoly relevant to the invention.

```
A device according to any preced-
ing claim , <wall /> further com-
prising illumination means
( 460 ) <wall /> for illuminating
the eye of said user , <wall />
wherein said viewpoint detecting       (4)
means <wall /> is adapted to de-
tect said viewpoint <wall /> by
receiving the light emitted by
said illumination means <wall />
and reflected by the surface of
said eye .
```

The claim presented in example (4) has more than 50 tokens and it is by no means one of the longest claims. Such sentences represent a problem in MT due to the complexity involved in translating them. In order to address this problem, we used the resource-light marker-based chunker (Gough and Way, 2004) from MaTrEx to split each input sentence sent for translation into smaller, more translatable chunks. The chunker employs a set of closed-class (or 'marker') words such as determiners, prepositions, conjunctions, pronouns, etc. to identify the points at which the sentence should be segmented. We adapted the algorithm and placed some additional constraints on the chunker to avoid over-segmentation of the input as this would be counterproductive. The chunks were converted into decoding zones sepa-

rated by the "<wall />" mark-up as shown in example (4). Once translated, the segments were recombined to produce a single output sentence.

3.2 Adaptation to the IPC System

Patents are classified using an international taxonomy – the International Patent Classification[3] system (IPC) – created by the World Intellectual Property Organisation. This allows us to consider the possibility of training separate MT systems for each patent (sub-) domain. There are 8 main categories (A—H) on the top level of the IPC taxonomy. In Table 2, we present these 8 patent domains along with the distribution of our MAREC corpus across each one.

IPC Domain	Sentence pairs	English tokens	French tokens
A (Human necessities)	1.99	65	74
B (Performing Operations)	1.92	71	79
C (Chemistry)	2.29	70	79
D (Textiles; Papers)	0.19	6	7
E (Fixed constructions)	0.31	11	13
F (Mechanical Engineering)	0.77	29	33
G (Physics)	2.04	68	78
H (Electricity)	1.83	63	72
Total	**11.39**	**387**	**438**

Table 2 Domain distribution of the sentence pairs and the number of tokens in the English–French parallel corpus (millions)

3.3 Experiments

In our previous work on patent domain adaptations for English—Portuguese (Tinsley, et al. 2010), the data was very unevenly distributed across the IPC and thus the results were not very definitive. However, having the patent data distributed among more evenly here, as shown in Table 2, we have the opportunity to better test whether combining multi-domain MT models might improve the overall system accuracy, as has been suggested (Haque et al. 2009; Banerjee et al., 2010).

In order to test this, we selected the patent domains containing close to, or more than 2 million sentence pairs: A, B, C, G and H. For each of these domains, we had a test set (and a development set) comprising 1,000 held out sentences, and we built four systems with different combinations of "in-domain" data and "general" data from the other domains.

These four system configurations comprised language models and translation models trained on the aforementioned in-domain and general data. For example, on the test data for the IPC C domain (Chemistry), the following four translation systems were evaluated: (i) one that has both the translation model (including lexical and reordering models) and the language model trained on domain C data *only* – "in-domain" TM and LM; (ii) a second one that has only the translation model trained on the domain data – "in-domain" TM and "general" LM training on *all* available data; (iii) a third one that has the translation model trained on all available data and the language model trained on in-domain data only – "general" TM and "in-domain" LM; and (iv) the baseline system that has the translation and the language models trained on all available data – "general" TM and "general" LM.

The results of these experiments are shown in Table 3 for English to French in terms of BLEU (Papineni et al., 2002) and METEOR-NEXT (Denkowski and Lavie, 2010). METEOR-NEXT uses the modules for exact matches, stemming and paraphrasing.

Test set domain	In-domain TM, in-domain LM	In-domain TM, general LM	General TM, in-domain LM	General TM, general LM
A	56.81 / 65.52	**57.18 / 65.81**	55.59 / 64.41	56.21 / 65.45
B	55.75 / 65.54	**56.31 / 65.90**	54.59 / 64.45	55.57 / 65.76
C	59.73 / 68.52	59.93 / 68.58	58.96 / 67.98	**60.9 / 69.18**
G	54.97 / 65.61	**55.18 / 65.73**	54.58 / 64.90	54.74 / 65.32
H	55.30 / 65.50	**55.76 / 65.83**	54.47 / 64.85	55.18 / 65.61

Table 3 BLEU / METEOR-NEXT scores for En-to-Fr MT systems with different in-domain and general domain configurations

The findings here show that the systems with in-domain translation models and general language models perform better than the baseline in four of the five patent domains taken into con-

[3] http://www.wipo.int/classifications/ipc/

sideration.[4] Similar results were achieved from French to English.

As we suggested in Tinsley et al. (2010), these findings are likely due to the nature of the training data found in domain C; that is to say, frequent long-winded chemical formulae, complex compounds, etc. that are unlikely to be useful when translating more general text. Omitting this data from the in-domain translation models when evaluating on domains A, B, G, and H therefore gives rise to improved results. On the contrary, when translating more natural language that may occur in the test data of domain C, the additional data from the other domains comes in handy and thus we see better results when using a general translation model.

4 Comparative Evaluation

In order to approximate the relative performance of our patent translation system, we performed an automatic comparative evaluation against two commercial systems: Google Translate[5] and Systran[6]. For PLuTO, we used the system configuration which performed best in the evaluations presented previously: in-domain translation model and general language model.

The evaluation was carried on 5,000 sentence pairs comprising a combination of all of the test sets (A, B, C, G, H) shown in Table 3. Evaluation scores for the PLuTO system were calculated over the output from the 5 domain-specific systems as a pseudo system combination as opposed to averaging over the original set of scores. The full set of results from both English—French and French—English are given below in Table 4 and Table 5.

English–French	BLEU	METEOR
PLuTO	56.95	66.32
Google	42.67	57.00
Systran	31.62	50.12

Table 4 BLEU / METEOR-NEXT scores for the English–French MT systems

French–English	BLEU	METEOR
PLuTO	56.92	67.90
Google	42.52	59.65
Systran	28.90	53.67

Table 5 BLEU / METEOR-NEXT scores for the French--English MT systems

We see significantly higher translation performance from the PLuTO system compared to the Google and Systran systems. Additionally, the domain-adapted PLuTO systems show an improvement of 0.6-0.7 absolute BLEU points and 1 METEOR-NEXT point over the general domain PLuTO MT systems (Table 3).

In the near future, as a deliverable requirement of the PLuTO project, we intend to publish a comprehensive manual evaluation of our translation engines, including a comparative human evaluation against the two systems employed here.

5 Conclusions

In this paper we have presented the most recent work carried out on MT for patents in the PLuTO project. We described the updated architecture of the system and a number of methods for adapting MT to the patent domain. We demonstrated improvements in translation accuracy by exploiting combinations in in-domain and general data as relates to the IPC system and showed PLuTO MT quality to improve upon that of Google and Systran. Additionally, we presented two techniques we employed to allow our engines to better handle some of the particular characteristics of patent documents.

Acknowledgments

The PLuTO Project has received generous 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 no. 250416.

References

Almaghout, Hala, Jie Jiang, and Andy Way. 2010. The DCU machine translation systems for IWSLT 2010. In Proceedings of the 7th International Workshop on Spoken Language Translation Paris, France, pp.37—44

Banerjee, Pratyush, Jinhua Du, Baoli Li, Sudip Naskar, Andy Way and Josef Van Genabith. 2010. Combining Multi-Domain Statistical Machine Translation Models using Automatic Classifiers. In AMTA 2010: The Ninth Conference of the Associ-

[4] We have not tested these results for statistical significance. In the near future, we intend to publish a large scale manual evaluation of the translation results which will serve as the definitive barometer.

[5] http://translate.google.com/

[6] The Systran system was used out of the box and not tuned to specifically to patents.

ation for Machine Translation in the Americas, Proceedings, Denver, CO., pp.141--150.

Dandapat, Sandipan, Mikel Forcada, Declan Groves, Sergio Penkale, John Tinsley and Andy Way. 2010. OpenMaTrEx: A free/open-source marker-driven example-based machine translation system. In Advances in Natural Language Processing, 7th International Conference on Natural Language Processing, IceTaL 2010, Reykjavik, Iceland, LNAI Vol. 6233, Springer, pp.121--126.

Denkowski, Michael and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages, *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR, 2010*

Gough, Nano, and Andy Way. 2004. Robust Large-Scale EBMT with Marker-Based Segmentation. In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04), pages 95–104, Baltimore, MD

Haque, Rejwanul, Sudip Kumar Naskar, Josef van Genabith and Andy Way. 2009. Experiments on Domain Adaptation for English-Hindi SMT. In Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation Hong Kong, pp.670–677

Hassan, Hany, Khalil Sima'an, and Andy Way. 2007. Supertagged Phrase-based Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), pp. 288–295, Prague, Czech Republic

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (ACL-2007), pages 177-180, Prague, Czech Republic

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'07), pages 304–311, Prague, Czech Republic

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics* (ACL-2002), pages 295-302, Philadelphia, PA, USA

Okita, Tsuyoshi, Jie Jiang, Rejwanul Haque, Hala Al-Maghout, Jinhua Du, Sudip Naskar and Andy Way. 2010. MaTrEx: the DCU MT System for NTCIR-8. In Proceedings of NTCIR-8, Tokyo, Japan, pp.377-383

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL), Philadelphia, July 2002, pp. 311-318

Penkale, Sergio, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. For-cada, 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.

Srivastava, Ankit, Rejwanul Haque, Sudip Naskar and Andy Way. 2008. MaTrEx: the DCU MT System for ICON 2008. In Proceedings of the NLP Tools Contest: Statistical Machine Translation (English to Hindi), 6th International Conference on Natural Language Processing, Pune, India

Stroppa, Nicolas, and Andy 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.

Stroppa, Nicolas, Declan Groves, Andy Way, and Kepa Sarasola. 2006. Example-based machine translation of the Basque language. In *Proceedings of AMTA 2006*, pages 232-241

Tinsley, John, Yanjun Ma, Sylvia Ozdowska and Andy Way. 2008. MaTrEx: the DCU MT System for WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation, ACL 2008*, Columbus, OH.

Tinsley, John, and Andy Way. 2009. *Automatically-Generated Parallel Treebanks and their Exploitability in Phrase-Based Statistical Machine Translation*. In Machine Translation 34(1):1—22.

Tinsley, John, Andy Way, and Páraic Sheridan. 2010. *PLuTO: MT for Online Patent Translation*. In Proceedings of the 9th Conference of the Association for Machine Translation in the Americas. Denver, CO, USA.