IPTranslator: Facilitating Patent Search with Machine Translation

John Tinsley
CNGL
School of Computing
Dublin City University, Ireland
john@iptranslator.com

Alexandru Ceausu
CNGL
School of Computing
Dublin City University, Ireland
aceausu@computing.dcu.ie

Jian Zhang
CNGL
School of Computing
Dublin City University, Ireland
jzhang@computing.dcu.ie

Heidi Depraetere
CrossLang
Woodrow Wilsonplein 7
9000 Gent - Belgium
heidi.depraetere@crosslang.com

Joeri Van de Walle
CrossLang
Woodrow Wilsonplein 7
9000 Gent - Belgium
joeri.vandewalle@crosslang.com

Abstract

Intellectual Property professionals frequently need to carry out patent searches for a variety of reasons. During a typical search, they will retrieve approximately 30% of their results in a foreign language. The machine translation (MT) options currently available to patent searchers for these foreign-language patents vary in their quality, consistency, and general level of service. In this article, we introduce IPTranslator; an MT web service designed to cater for the needs of patent searchers. At the core of IPTranslator is a set of MT systems developed specifically for translating patent text. We describe the challenges faced in adapting MT technology to such a complex domain, and how the systems were evaluated to ensure that the quality was fit for purpose. Finally, we present the framework through which the IPTranslator service is delivered to users, and the value-adding features which address many of the issues with existing solutions.

1 Introduction

The accessibility of online machine translation (MT) tools such as Google Translate[1] and Bing Translator[2] have conditioned people to viewing MT as a free service. When considering the commercial opportunities for pure MT (i.e. not as part of a CAT[3] tool or language services offering), vendors must think outside the box in terms of what value they can add to the translation service in order to make it an attractive proposition for potential customers. As the free MT services continue to improve, translation quality becomes a harder sell, so the added-value may need to be of a more customised nature, for example, a feature set tailored to a specific type of user.

In this article we describe the development of IPTranslator – an MT framework for translating patent documents – and how this domain-specific translation service has been adapted to be incorporated into the daily workflow of a particular set of users; in this case, patent searchers.

We begin by introducing the role of the patent searcher, how they operate, and the scenarios which give rise to the need for MT, including the level of translation quality they require.

Following this, we discuss the difficulties facing MT when it comes to the extreme linguistic complexity of patent text and how we go about tackling these problems in our systems. Evaluation is obviously a key step in the development any translation system, but it takes on an even more pivotal role when the translations are intended to be used for a specific task. To that end, we describe the measures we have employed not only to

[1] http://translate.google.com
[2] http://www.bing.com/translator

assess the quality of our translations but also their suitability for the task of the patent searcher.

As IPTranslator is delivered to users as a web service, we also describe some of the challenges involved in deploying MT as an on-demand online service. Finally, we present some of the features implemented in IPTranslator which have been designed to add value to the MT output for patent searchers and to facilitate as seamless an integration as possible into their workflow.

2 Patent Searching

Intellectual Property (IP) professionals – patent searchers, patent attorneys, patent examiners – frequently need to carry out patent searches for a variety of reasons, for example, to judge whether a new idea or invention is patentable (patentability search), or to assess whether a competitor's patent infringes on an existing one (infringement search). Typically, the searcher will compile a set of 50-200 patents that are potentially relevant to their task based on a first assessment. Following this, they will take a much closer look at these patents, perhaps in collaboration with colleagues or legal representatives, to take a final decision (proceed with application, take legal action, etc.)

The IP professionals in question here are predominantly patent specialists who work for large multinationals that are actively developing new systems and processes. Common examples of such companies can be seen in the pharmaceutical industry (e.g. Bayer, GlaxoSmithKline, Pfizer, Novartis) and the computing and consumer electronics industries (e.g. Apple, Samsung, Microsoft). They work closely with the engineers and scientists "on the ground" in their organisation to identify innovations in addition to monitoring the IP landscape in relation to their employer's patent portfolio. Other types of individuals who carry out similar patent searches include legal professionals, such as patent attorneys, officers in university technology transfer offices, as well as individual inventors and entrepreneurs.

Because there is no single database of patents, searchers must use multiple different resources to try to find relevant documents and ensure that their search is as comprehensive as possible. These include searchable databases at national patent offices, e.g. the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), specialised agencies like the World Intellectual Property Organisation (WIPO), proprietary collections, e.g. Thomson Reuters (Derwent World Patent Index), Minesoft's PatBase, or Questel's Orbit service, and other resources such as general search engines and collections of non-patent literature, e.g. scientific articles.

2.1 Language Barriers

Language barriers are no excuse in the case of patent infringement. A company is still liable, for example, if their English language patent filed in the UK infringes on a Chinese patent. It is the responsibility of the company (including the IP specialists and legal counsel) to ensure that they have freedom to operate and to protect their own patents. This is done via the aforementioned patent searches. However, the problem lies in the fact that, during the compilation of the set of potentially relevant patents, around 30% of the results on average are in a foreign language.

Professional translation of patents is very expensive and is often not justifiable, particularly at this stage in the search process when the outcome might be to learn that the document was not relevant in the first place. Given this, searchers typically use machine translation to at least try to get the gist of a document and determine whether it is worth investing in a human translation. It is at this point the problems begin to mount up.

Looking at trends in patent filing over the last number of years, it is clear the need for patent translation is only going to increase; particularly for Asian languages. In Figure 1, we see that almost 50% of all patent applications filed worldwide in 2009 came from China, Japan, and South Korea. This represents an upward trend for those countries which, according to WIPO's 2012 yearly review (WIPO, 2012), continued to show an increase in applications from 2010 to 2011 of 33%, 21%, and 8% for China, Japan, and South Korea respectively (compared to an 8% and 6% increase for the USA and Germany respectively; two similarly IP active countries).

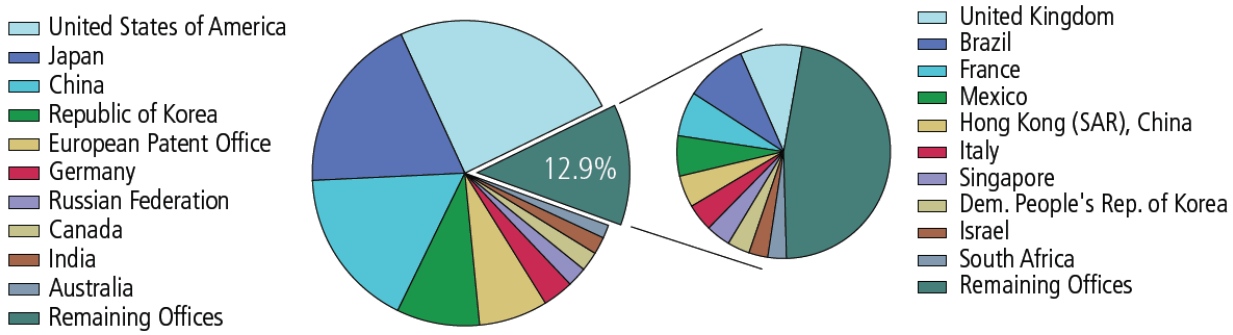


Figure 1 Proportion of patent applications at the top 20 IP offices in 2009 *(WIPO, 2011)*

2.2 Existing Solutions

The options currently available to patent searchers vary depending on where they are carrying out the search. The search tools mentioned in the previous section offer different levels of service, if they offer it at all.

For instance, many of them integrate with Google Translate in some way (EPO Espacenet, WIPO Patentscope, Questel Orbit), others have their developed their own systems for certain languages (Thomson Innovation, PatBase, Japanese Patent Office, Korean Patent Office), while some have no integrated service (German Patent Office) forcing the searcher to seek translation elsewhere.

The principal issue with these translation services is that the MT systems have typically been developed for general use, i.e. to translate any type of text. For example, the Google Translate system used to translate patents in the EPO's Espacenet service is the same Google Translate a user might use to translate a sports news website or a help forum for C++ programmers.[4] However, research has widely demonstrated that systems which are specifically adapted to translate text in a particular domain, e.g. patents, outperform general purpose and out-of-domain systems (Koehn and Schroeder, 2007; Bertoldi and Federico, 2009; Banerjee et al., 2011).

Another problem with the translation options available to patent searchers is the manner in which the service is delivered. Given that the different search tools provide different translation options and that they are integrated in different ways depending on the tool (via API, widget, Google Chrome plugin), from the patent searcher's perspective there is an inconsistency of translation quality and user experience, as well as the lack of ability to interact in any way with the translation process. Furthermore, users will often visit alternative translation services not integrated with their preferred search tools in order to view a "second opinion".

Finally, in addition to translation during the search process itself, the patent searcher may have documents retrieved from others sources (colleagues, non-patent literature via the web) which require translation. These documents come in different formats, e.g. MS Word, PDF, and in these instances there are even fewer options for translation.

3 Our Solution: IPTranslator

The product we have developed – IPTranslator[5] – aims to address the key issues mentioned above, namely: the lack of a single resource for patent specific MT; the inconsistency of translation quality; the inconsistency of user experience; the inability to influence the translation process; and the inability to act on the translation output (edit, extract meaning, share, etc.).

To date we have developed domain-specific MT systems for six language pairs – English to/from French, German, Spanish, Portuguese, Chinese, and Japanese – and built a software platform which exposes them in a manner that is fully integrated into the workflow of the patent searcher.

In the following, we discuss the challenges faced in developing such MT systems, how we went about ensuring that they were fit for purpose,

[4] This is despite marketing to the effect that Google has developed a standalone "Patent Translate" service based on a collaborative agreement signed with the EPO.

[5] http://www.iptranslator.com

Rule description	In patents
Sentences should be short, i.e. around 25-30 words, to reduce decoding complexity.	Patent sentences are usually very long, even containing up to 500 words (see Appendix I for an example sentence containing 152 words).
Spelling should be correct to avoid a high out-of-vocabulary (OOV) rate.	Many patents are only available in digital format by means of OCR[6] which introduces orthographic errors in the source.
Sentences should be grammatically complete and not written in telegraphic or nominal style.	Patents often make use of nominal or telegraphic style, especially in titles, e.g. "Handheld Processing Device Including Medical Applications For Minimally And Non Invasive Glucose Measurements".

Table 1 Examples of controlled language rules and how they are broken in patents

and how we ultimately brought them to the end-users.

3.1 Patent Translation is Hard

Patents are highly technical in nature make substantial use of well-established conventions when authored. For instance, patent claims typically contain a preamble, a 'transitional' phrase, and a description of the invention, all of which are expressed in a single sentence.

As a consequence of this, patents are often a rich source of extreme syntactic complexity making extensive use of such features as nominal style, relative clauses, and neologisms (Rossi and Wiggins, 2012). Additionally, patents can contain terms and formulations with a high level of technicality, such as chemical compounds, intended for the expert reader. These features present a real challenge to the developers of MT systems.

The field of Controlled Language prescribes a set of recommendations for the authoring of text in order to reduce ambiguity and complexity and to make documents more intelligible to humans and also more amenable to being processed automatically (Roturier, 2006; O'Brien and Roturier, 2007). These recommendations include a set of rules designed specifically to increase "machine translatability" of a body of text. Patents exhibit a large number of characteristics that are particularly non-compliant with these rules.

Table 1 presents an excerpt adapted from Wiggins (2012) which illustrates how patents specifically do not comply with some of these rules.

[6] Optical Character Recognition

3.2 Our Approach

The first step in the process of building MT systems for a specific task is to acquire sufficient *in-domain* data for training. All of our systems have been training predominantly on large parallel patent corpora. As patents deal with all areas of human knowledge, it is clear that we can further divide patents into various sub-domains, e.g. chemistry, engineering, physics, etc. We previously investigated a number of different approaches to adapting the MT systems to these various sub-domains. The outcome of these experiments, described in full in Ceausu et al., 2011) showed that the performance of the various sub-systems was heavily dependent on the distribution of training data across the sub-domains.

Following the development of the patent-specific models, we implemented a number of additional procedures to deal with some of the aforementioned characteristics of patents which can cause issues for MT. For example, a sentence splitting module has been developed to break input sentences into smaller more translatable chunks of text and also to improve the efficiency of the service by making the most of our parallel computing resources. Specific sequences which can pose challenges, such as chemical formulae and references to images, are isolated prior to translation using an approach based on named-entity recognition in order to handle them separately.

In addition to the procedures designed to account for the specifics of patents, we also employed a number of language-specific processes to deal with factors like segmentation for Chinese, compound splitting and joining for German, as well as long distance reordering the many of the language pairs, including Japanese and German.

4 Evaluation

Our goal in designing an evaluation framework for the patent translation systems was to be able to answer two questions: 1) is the translation quality good? And 2) is the translation quality good *enough*?

Before deploying a particular system in our application, we try to ensure that the answer to the first question is positive. To do this we evaluate all systems using both automatic and human measures. Iterative developments are made to the systems and assessed using automatic metrics until the results have stabilised above a sufficient predetermined threshold. Only at this stage are human evaluators employed to provide judgements, error analyses, and perform benchmarking against competing systems.

4.1 Automatic Evaluation

For the automatic evaluation, we used a test set of 8,000 sentences pairs comprising 1,000 sentences from each of the top 8 sub-domains of the International Patent Classification (IPC) system.[7] Scores were calculated using the BLEU[8] metric for each of the sub-domains in order to assess the relative performance in each area. We also calculated scores for Google Translate and Systran on the same test sets. The results of these evaluations are shown in Table 2.

Sub-domain	IPTranslator	Google	Systran
All	56.28	43.32	32.92
A (Human necessities)	56.21	42.67	31.62
B (Operations)	55.57	44.58	33.82
C (Chemistry)	60.90	45.92	31.72
D (Textiles)	58.00	44.8	33.09
E (Fixed constructions)	52.64	41.93	32.30
F (Mechanical engineering)	56.69	45.34	35.00
G (Physics)	54.74	40.24	32.69
H (Electricity)	55.18	40.96	32.40

Table 2 Automatic scores for French to English

[7] http://www.wipo.int/ipcpub/
[8] http://en.wikipedia.org/wiki/BLEU

Overall we can see that our system achieved relatively high scores, particularly for the chemistry sub-domain which is one of areas in which patenting is most active. It performed significantly better than the other MT systems, showing a 30% relative improvement over Google Translate and a 71% improvement over Systran. The key question at this stage is whether these results correlate with expert human opinion.

4.2 Human Evaluation

The human evaluations were focused on the linguistic quality of the translations and how they compared to the output of other MT systems. To assess the adequacy of the translations in linguistic terms, informants were asked to evaluate the translation quality of each individual translated sentence in the human evaluation set by giving it a score from 1 (Very Poor) to 5 (Excellent). The results from this evaluation for French to English translation, shown below in Figure 2, were positive with the three evaluators giving an average score of 3.88 to the MT output based on a set of 800 sentences.

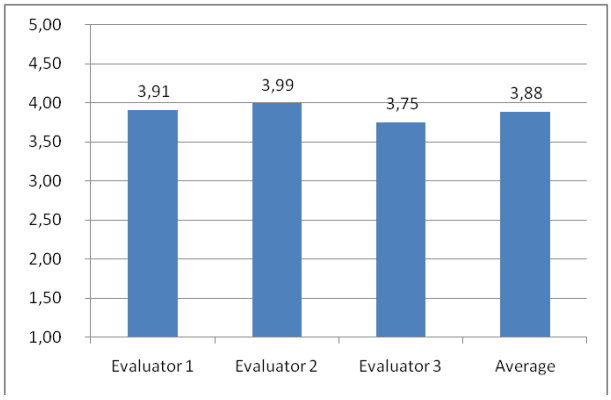


Figure 2 Human ranking evaluation results for French-English

Following this, in order to benchmark our system in relation to competitors, informants were asked to compare the output of three different MT systems – IPTranslator, Google Translate, and Systran – and rank them in order of perceived quality. This evaluation was blind in that informants were not aware of the origin of the translation output. The outcome of this evaluation is shown in Figure 4. We see the evaluators typically preferred the IPTranslator output, ranking it first 67% of the

time and as the worst translation only 6% of the time. In cases where the evaluator could not choose between translations, they were given the same ranking.

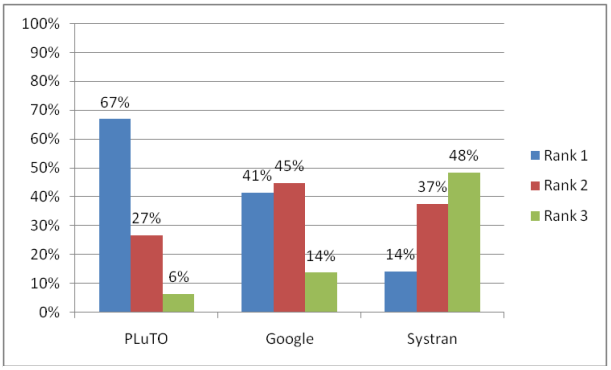


Figure 3 Human benchmarking evaluation results for French-English

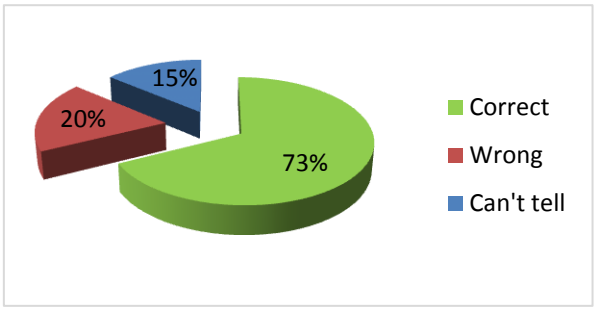
4.3 Usability and Utility to Patent Searchers

Having satisfied ourselves that the overall translation quality was acceptable, the next step was to design an evaluation to assess whether the translations actually meet the needs of potential end users. This type of usability evaluation goes beyond the classical approach to MT evaluation in that it is user centred and takes into account use cases of translated text.

We designed an experiment to simulate the task of the patent searcher in which, given a particular invention and a set of machine translated patents (hypothetically retrieved during a patent search), they must judge whether the patents are relevant or not to the invention. In conjunction with members of the Dutch Patent Information User Group (WON)[9], a set of French patents were carefully selected to ensure that 50% were relevant and the other 50% not relevant. The documents were then machine translated into English and group of patent searchers were asked to make the judgments.

In total, we had 11 participants take part in this experiment using a set of 20 patents across two sub-domains (chemistry and mechanical engineering: 10 patents and 1 invention reference per sub-domain). The results are summarised in **Error! Reference source not found.**.

These results were broadly positive in that the patent searchers were able to make the correct



judgement 73% of the time based on the MT output. Digging a little deeper into these results we learned that in 90% of the cases where the searcher could not make a judgement, they attributed it to a lack of detail in the description of the invention as opposed to poor translation quality.

Figure 5 Results of the usability experiment for French-English

As an addendum to this experiment, we also asked the searchers to give a snap judgement on the translation quality of the documents they read. Figure 6 below shows that the translations were perceived well by the searchers on the whole, with only 12% of documents ranked poor or worse.

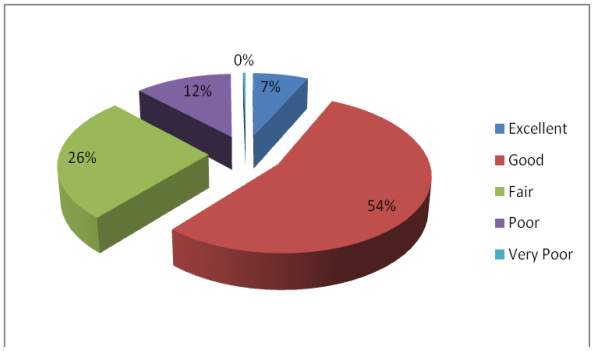


Figure 6 Results of the snap judgement by patents searchers on translation quality

While the results presented here are for the French—English language pair, we have carried it out for other pairs with similar results. Our ultimate goal is to replicate this experiment for each language pair before we bring it out of beta.

5 MT as a service

As we described in Tinsley et al. (2010), delivering a machine translation web service in which MT systems must be on-demand for a number of language pairs 24/7, is non-trivial. The challenge is to deliver translations in a timely manner within the limits of our infrastructure. Without the seemingly

[9] http://www.won-nl.org

unlimited computational resources of the likes of Google or Microsoft at our disposal, we have to be clever in our deployment.

There are three main factors we have to take into consideration: translation quality, translation speed, and the computational resources required. Essentially, we have to find an optimal trade-off between these three factors depending on the task.

In order to establish the optimal configuration, we carried out extensive testing prior to deployment, described previously in Tinsley et al. (2010), in which we trained a number of systems with different size language models and phrase tables pruned with different degrees of aggression. The results of these experiments are illustrated in Figure 7 where each bubble represents a particular MT system configuration and the larger the bubble, the larger its memory footprints (memory values in Gb are given beside each bubble).

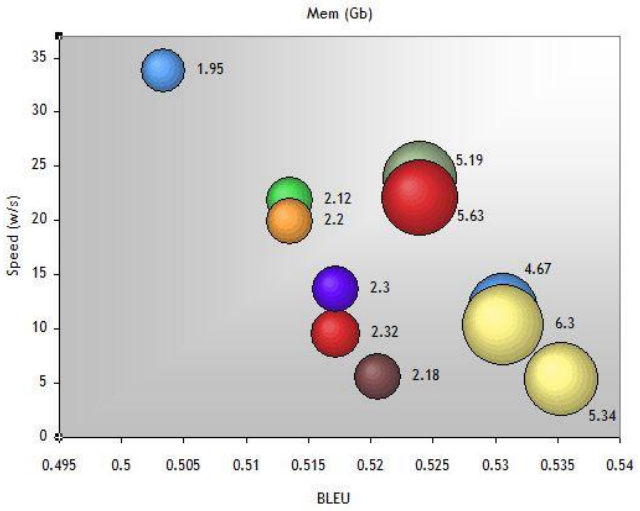


Figure 7 Balance between quality (BLEU), speed (words per second), and memory consumption (Gb). Each bubbles represents an MT system and the size of the bubble corresponds to its memory footprint

The first point to note is that selecting the "best" system is not that straightforward. Generally, we can see that the bigger the translation models the better the quality, the more memory is needed, and the slower the translation is.

In our case, feedback from users suggests that some sacrifice in speed is acceptable for an increase in translation quality which is one element where we have some room to manoeuver. Additional existing approaches to storing and retrieving data from the models in an efficient manner, as well as novel methods for pruning phrase tables, give developers further options when it comes to finding the optimal solution.

6 Adding Vale to Translation

Users of MT in the patent search space have been conditioned to viewing automatic translation as a free service; typically integrated as an 'add-on' to the search tools we introduced in section 2. Therefore, in order for us to build a viable business model around the MT systems we have developed, we must somehow add value beyond just the translation quality (which, anecdotally, can be difficult to demonstrate).

To do this, we carried out an audit of MT services in the patent search field to identify areas in which they fall short. For each issue, we developed a specific solution and incorporated this into the user interface. Some examples are given in the following:

Issue: The different levels of translation service offered by the various search tools causes patent searchers to have to jump from site to site to find translations. This leads to a general inconsistency in the translation experience for end users.
Solution: We have developed a browser plugin tool through which users can access IPTranslator on-the-fly, no matter where they carry out their patent searches.

Issue: Patents are full of highly technical terms and neologisms which gives rise to a high OOV rate. As patent searchers typically work in the same technical area for periods of time, they see repeated instances of untranslated words and errors.
Solution: We allow users to add their own terminology and to correct mistakes (including untranslated words) in the translation output. These edits are remembered for a particular user and applied in subsequent translations.

Issue: While newer patents exist in fully digital format, many older patents only exist in PDF format from scanned documents. There are limited (none integrated) options for translating these.
Solution: We provide a facility for PDF translation which allows the user to upload documents for translation. For scanned PDFs, we used OCR tech-

nology, trained on patent data, coupled with noise filtering, to increase the quality.

Issue: The whole concept of IPTranslator is to improve the efficiency of the searcher. Patents can be large documents and searchers need to be able to navigate them quickly and extract relevant information. While features to support this are available for original language documents, there are no such solutions for translated documents.
Solution: We do segment highlighting (on a phrase and sentence level) to show the corresponding segments between the original patent and its translation. Bilingual keywords are also extracted to summarise the document and provide additional information for subsequent cross-language searches.

Issue: Patent searching within an organisation is often a collaborative process where searchers need to be able to share their output with colleagues. Again, this option is not always available for translated documents.
Solution: We allow users to download translations and we intend to develop a facility through which documents can be shared amongst authorized parties.

7 Summary

We have described IPTranslator which delivers "machine translation as a service" to patent search professionals. We have presented some of the challenges posed by patents for MT and demonstrated that systems which have been developed to specifically cater for a particular type of text can yield better results than the best general domain systems.

Task-based evaluation is important when building MT systems for a particular purpose and we have presented the design and execution of such an evaluation to ensure not only that our translation quality is good but also that it is fit for purpose.

In addition to the translation technology itself, we have discussed some of the obstacles faced when delivering MT as an on-demand service including the trade-off that needs to be made between translation speed and quality given the available computational resources. Ultimately, the decision will be based upon the needs of the user and whether there is any slack to be found in one of the key factors.

Finally, we presented some of the value-adding features of IPTranslator designed to enhance the translation experience and plug some of the gaps in existing services.

References

Banerjee, P., S. K. Naskar, J. Roturier, A. Way and J. Van Genabith. (2011) *Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component Level Mixture Modelling*. In the Proceedings of the Machine Translation Summit XIII. Xiamen, China. pp 285-292.

Bertoldi, N. and Federico, M. (2009) *Domain Adaptation for Statistical Machine Translation with Monolingual Resources.* In Proceedings of the Fourth Workshop on Statistical Machine Translation, StatMT '09 Morristown, NJ, pp 182–189

Ceausu, Alezandru, John Tinsley, Jian Zhang, and Andy Way. (2011) *Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project*. In Proceedings of EAMT 2011: The 15th Annual Conference of the European Association for Machine Translation. Leuven, Belgium

Koehn, P. and J. Schroeder. (2007) *Experiments in domain adaptation for statistical machine translation.* In Proceedings of the Second Workshop on Statistical Machine Translation Prague, Czech Republic, pp.224–227

O'Brien, S. and J. Roturier (2007) How Portable are Controlled Language Rules? A Comparison of Two Empirical MT Studies. Proceedings of MT Summit XI. Copenhagen, Denmark. pp. 345-352.

Rossi, Laura, and Dion Wiggins (2012) *Applicability and application of Machine Translation quality metrics in the patent field.* In World Patent Information. Eds. Michael Blackman (to appear)

Roturier, J. (2006) *An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine-Translated Technical Documentation for French and German Users*. Unpublished PhD thesis, Dublin City University, Ireland

Tinsley, John, 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.

WIPO (2011) *WIPO IP Facts and Figures 2011*. WIPO Publication No. 943(E). ISBN 978-92-805-2111-5

WIPO (2012) *PCT Yearly Review: The International Patent System.* WIPO Economics and Statistics Series. WIPO Publication No. 901E/2012. ISBM 978-92-805-2238-9\

Appendix I

The reconstitution container device of claim 1 further for drying liquid medication wherein the outer shape of the barrel component is selected so that before assembly of the plug component with the barrel component, barrel component may be mounted in a support device with the winding mixing channel in an open channel configuration with the open portion of the channel facing upward and with liquid medication residing in the open cavity for the purpose of undergoing a drying process and wherein the distal end of the plug component has a flattened shape such that after the liquid medication has been dried to a powder and the plug component is engaged with the barrel component, the flat distal surface pushes the powder fully into the mixing channel and closes the top of the mixing channel wherein the reconstitution container device is useful for drying liquid medication, storing dried medication, and reconstituting dried medication.

I Example of a single sentence from a patent claim containing 152 words