Tapta: A user-driven translation system for patent documents based on domain-aware Statistical Machine Translation

Bruno Pouliquen **Christophe Mazenc** **Aldo Iorio**
World Intellectual Property Organization
34, chemin des Colombettes
CH-1211 Geneva 20
Bruno.Pouliquen@wipo.int, Christophe.Mazenc@wipo.int, Aldo.Iorio@wipo.int

Abstract

This paper presents a study conducted in the course of implementing a project in the World Intellectual Property Organization (WIPO) on assisted translation of patent abstracts and titles from English to French. The tool (called 'Tapta') is trained on an extensive corpus of manually translated patents. These patents are classified, each class belonging to one of the 32 predefined domains. The trained Statistical Machine Translation (SMT) tool uses this additional information to propose more accurate translations according to the context. The performance of the SMT system was shown to be above the current state of the art, but, in order to produce an acceptable translation, a human has to supervise the process. Therefore, a graphical user interface was built in which the translator drives the automatic translation process. A significant experiment with human operators was conducted within WIPO, the output was judged to be successful and a project to use Tapta in production is now under discussion.

1 Introduction

This paper describes a tool called Tapta[1] aimed at facilitating the work of translators in the World Intellectual Property Organization (WIPO) which is a specialized agency of the United Nations[2]. One of the activities of WIPO is to administer the Patent Cooperation Treaty (PCT) which offers patent applicants an advantageous route for obtaining patent protection in multiple countries. For publication purposes, WIPO translates the title and the abstract of each received patent document (called a PCT application) from its filing language into both the English and French languages. The number of published documents per year has been steadily increasing in the last two decades (from 16,000 in 1990, 80,000 in 2000 to 150,000 in 2010). Originally only performed by WIPO translators, a large share of these translations (more than 80%) is now outsourced to private companies and we are looking at the latest achievements in Computer Aided Translation (CAT) tools and Machine Translation (MT) to optimize costs and improve quality. High quality of translation is required, WIPO conducts a quality check. PCT statistics are published in WIPO (2010).

We have an extensive parallel corpus of manually translated patent documents collected over time, especially for the language pair English-French (1.8 million documents). A test was launched: a statistical machine translation (SMT) system was trained using our corpus and a tool to help translation of patent titles and abstract from English to French provided. WIPO's expectations in term of quality of translation are extremely high, therefore using solely the results of an SMT was out of question. To meet these requirements human judgment was needed in addition to the computer.

The system presented in this paper differs from other systems in the sense that it relies on a technical-domain-aware SMT using *factors* (i.e. a SMT taking into account as an additional parameter the technical domain each text segment belongs to), and it relies on the user to drive the segmentation of the text and subsequently to select the best proposal.

2 State of the art

There is a growing demand for patent translations (Tsai 2008). Automatic translation of patent documents can help in a few cases: it offers the user the possibility to understand a patent written in a language he does not know, it provides trans-

[1] Translation Assistant for Patent Titles and Abstracts
[2] "WIPO is dedicated to developing a balanced and accessible international intellectual property (IP) system, which rewards creativity, stimulates innovation and contributes to economic development while safeguarding the public interest" http://www.wipo.int/about-wipo

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 5–12
Leuven, Belgium, May 2011

lators with a first draft translation to post-edit and, finally, it can be used as an interactive tool to automate part of the translation process.

Automatic translation of patents is catching the attention of international actors as a means to overcome the language barrier, for example: Patent Machine Translation Task at NTCIR-9[3], the European project Pluto (Tinsley et al. 2010) and the collaboration between the European Patent Office and Google translate[4], etc.

Various techniques can be used in Machine Translation (Koehn 2010): rule-based systems, example based translation, statistical machine translation and hybrid systems.

In addition to machine translation tools widespread use of computer aided translation (CAT) tools to help translators in their daily work must be mentioned (provides quality assurance, glossaries, terminology, translation memories etc.).

An international organization like ours has 10 working languages (WIPO 2010), which means that, if such an organization wanted a translation tool in all language pair combinations, it would require 90 translation engines. Therefore a rule-based translation system is not affordable because of the burden to enter and maintain all rules and dictionaries for all language pairs. A data-driven approach is usually more suitable when a big parallel corpus exists (not to mention the human resources required). Among the two data-driven approaches, the example-based approach alone is not adequate in the patent domain as there are not enough repetitions in patent documents. For these reasons, the statistical approach (SMT) is more and more considered in the field of patent translation. Hybrid systems are still an alternative when one already has access to multiple engines (Thurmair 2009, Tinsley et al. 2010).

The most commonly used SMT system: *Moses* (Koehn et al. 2007) is highly configurable, available for free and is the only system, to our knowledge, that allows *factored translation models* (Koehn & Hoang, 2007). Factored translation models allow the enrichment of input texts tokens (words) with additional information (lemma, grammatical information, etc.), the decoder can then use all or parts of these *factors* to improve translation. Among the machine translation drawbacks, translation ambiguity is an important issue, an English term like 'automatic translation' will generally be translated as '*traduction automatique'* in French, but in a domain like mechanical engineering it would be translated as '*translation automatique*'. Previous experiments using two domains (encoded as a *factor*) were achieved with reasonable success by Koehn and Schroeder (2007) and Nihues and Waibel (2007). It will be explained, in section 3.2, how these *factors* were generalized to include the domain in the translation model.

When the "Fully Automatic High-Quality Machine Translation" goal is unreachable (such as in the WIPO context) humans need to participate in the translation process. Usually human intervention is limited to post-editing (correct errors produced by an automatic translation system and submit a final high-quality translation). This approach is usually the one adopted by most of the CAT systems. However, SMT systems are not limited to one translation per document, which is one drawback of usual post-editing CAT tools. On the contrary, SMT systems have lots of translation possibilities for each chunk of text. Recent work investigated the inclusion of human judgment at each step of the translation process: (a) *Caitra* system which shows the user various "paths" among all translations of a sentence (Koehn 2009), (b) *TransType* which is keyboard-driven by the user to select the best translation after each entered word (Macklovitch et al. 2005, Macklovitch 2006). Our system asks the user not only to select among possible translations, but also to drive the segmentation of the input text.

3 Training the SMT

3.1 The English-French parallel corpus

Our SMT is trained with the parallel corpus extracted from previously translated PCT applications (1,800,788 documents). Each translation contains a title and an abstract in both languages.

Each document is categorized manually in one or more IPC class[5]. A simple algorithm was built that determines the "domain(s)" of the document using the IPC class. Each document belongs to one of the 32 pre-defined domains[6] (maximum: 3 domains per document).

As mentioned in the introduction, WIPO runs a quality check (QC) on a sampling set of trans-

[3] This task proposes to participant to train patent machine translation tools on the same parallel corpus in English, Japanese and Chinese, and then compare the various techniques and results, see http://ntcir.nii.ac.jp/PatentMT

[4] See http://www.epo.org/topics/news/2010/20101130.html

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

[6] The 32 domains are high level categories of patents (Transport, Medicine, Energy, Foods, Chemistry, etc.).

lations. Parallel documents extracted from these *QC*-documents are therefore of better quality. In addition to the QC, the fact that an application that was originally in French has better linguistic quality in French is used. These documents are given a special weight (see section 3.3).

Each application contains two different parts: a title and an abstract with respective specificities:

- Titles are often: short (on average 7.8 words), in uppercase, without conjugated verbs, sometimes without accents in French

- Abstracts are often: long (on average 113 words), in lowercase, with more than one sentence (3.38 on average), references in brackets and accentuated in French.

To obtain the best training corpus we apply the common following steps:

- filter out French abstracts without accents (which is the case only for old applications)
- re-accentuate French titles taking into account the accentuated abstract, filter out titles having ambiguous accentuation
- carry out sentence splitting of abstracts
- tokenize abstract sentences and the title
- use *Champollion* (Ma 2006) to align abstract sentences
- filter out badly aligned sentences
- filter out sentences having only one word or more than 80 words[7]
- filter out pairs of sentences where the ratio (number of English words/number of French words is more than 9)
- apply some regular expression replacement rules (delete xml tags, uniform accents, etc.)

Our own tokenization tool was developed (based on *Lucene* tokenizer (McCandless, 2010)) in order to split words according to the specific language used in patents (references, Greek letters, apostrophe). Our sentence splitter relies on sentence boundaries and list of abbreviations. When the sentence is too long, our tool splits also on segment boundaries (i.e. comma, semi-column, reference etc.), see Table 1 for an example.

A robust sentence aligner was also needed, as sentences and segments are not always exactly aligned segment by segment, therefore we adapted *Champollion* (Ma 2006) so that it further splits the sentences in sub-segments when the alignment is not good enough. *Champollion* needs to be fed with a bilingual lexicon, an IBM model 1 was first run on titles only to learn the lexicon from the data, then *Champollion* was fed with the obtained lexicon [8]. Unaligned segments and segments aligned with a very low lexical weight (usually errors in the translation, or strange alignments due for example to omissions of dots in one of the abstract) were filtered out. As a result 8’352’768 aligned pairs of segments (high quality) were obtained.

Table 1: Tokenization/alignment example with underlined segment boundaries (here the best alignment is achieved by using references, notice the last English segment which is aligned with two French segments)

a · pesticide · and/or · a · fertilizer · and · the · like · scattered · or · sprayed · onto · a · golf · course · (1) · , · together · with · most · of · the · sprayed · water · or · rain · water · , · are · permated · into · a · turf · (15A) · , · a · thick · sand · layer · (15) · having · high · water · permeability · , · whereby · only · the · rain · water · filtered · by · the · golf · course · is · discharged · to · the · outside	un · pesticide · et/ou · un · fertilisant · , · ou · un · produit · similaire · , · dispersé · ou · pulvérisé · sur · un · parcours · de · golf · (1) · , · avec · la · majeure · partie · de · l' · eau · pulvérisée · ou · de · l' · eau · de · pluie · , · qui · sont · imprégnés · dans · un · gazon · (15A) · , · une · épaisse · couche · de · sable · (15) · ayant · une · forte · perméabilité · à · l' · eau · . · Selon · ce · procédé · seule · l' · eau · de · pluie · filtrée · par · le · sol · du · parcours · de · golf · se · déverse · à · l' · extérieur

3.2 SMT training

Once the parallel corpus is available, the SMT training can be launched using the freely available tool *Moses*.

A few documents were retained as our development set in order to carry out the evaluation (it was decided to randomly select 102 segments out of 52 quality checked documents). The test set was selected as a random selection of the latest published applications (9521 segments out of 1390 documents).

One specific feature of our training in comparison to similar approaches is that the translation is dependent on the text *domain*. It was decided to encode the domain information using *Moses* decoder’s “factored model” facility (Koehn & Hoang, 2007).

Such an approach has already been used by Koehn & Schroeder (2007) and Nihues and Waibel (2007), however they used a binary factor information: *in-domain* versus *out-of-domain*

[7] This filter is maybe too aggressive but the word alignment quality will usually be poor on such big sentences.

[8] Note that the lexicon obtained after a full training on titles and abstracts is generally better than the one learned on titles only. To overcome this “chicken and egg” problem, we also include the previously learned lexicon. Experiments should be carried on to evaluate the number of loops necessary before the system converges.

corpus. In our case, factors are used to provide various translations according to our 32 domains[9]. Each word in the input sentence is "tagged" with the domain of the application in such a way that the decoder will give preference to translations in the same domain. As a result the translation of 'rope' in French will differ according to the domain (i.e. '*câble'* – a metal rope - in the *civil engineering* domain and '*corde'* – non-metal rope - in other domains).

A "domain-tag" is first attached to each word of the input text depending on its domain. But an evaluation showed that the domain translation was better when we did not attach specific domain to commonly used words (like 'a', 'the', 'whereis', … or even 'invention', 'process', 'main', 'small', etc.). Such a list was compiled automatically by comparing frequency of words across all domains. As a result, a factor is added (one letter identifying the domain of the text, or 'U' when undefined) to each word of the sentence. As example: the two following sentences have different factors attached: '*in|U the|U robotic|1 prosthesis|1 alignment|1 device|U provides|U automatic|1 translation|1 in|U two|U axes|1*' and '*a|U chinese|2 to|U english|2 automatic|2 translation|2 method|U*', the 'domain-aware' translation decoder will then be able to disambiguate the two meanings of '*automatic translation*' according to the domain.

Mgiza++ (Gao & Vogel, 2008) was used to align words in sentences. At this stage the domain information was not used. On a *linux* server (eight 2.5 Ghz cores) it ran for 5 days on our corpus. An additional step stores this information in a *Lucene* index so we can use it for our concordancer (see section 4.2).

Two language models are built using IRSTLM toolkit (Federico et al. 2008) with 5-grams (without domain information). One model is generated out of the set of French texts out of our parallel corpus (1.8 million), the other is built out of a bigger corpus of all French patent documents provided by WIPO (4 million) (without any document belonging to the test or development set).

3.3 Optimization

Attempts to optimize the performance of the system with various settings were carried out:

The generated phrase tables are huge (312 million entries) and contain phrases that are very unlikely to be seen in another context. Our corpus is big enough to require a minimum frequency for a phrase to be included in the phrase table. The "pruning" method as described in Johnson et al. (2007) was tried. The suggested parameters were retained i.e.: delete all phrases which occur only once in our training corpus and, for each phrase, only the first 30 translation candidates are kept. The "pruned" phrase tables now contained 102 million entries (33% of the original) and even more importantly the quality and speed of the translation improved as shown in Table 2.

In order to take into account our Quality Control, a separate set of aligned sentences was created belonging to applications having passed the QC or that were published in French (this step is run after *Mgiza*). New phrase tables were learned from this subset. The two phrase tables were then merged (i.e. the table from the subset and the larger table, after various experiments we ended up using 25% of the subset and 75% of the larger table). As shown in Table 2, the QC-improved model alone cannot be used as it is too small.

4 Translating / graphical user interface

4.1 Server configuration

The *Moses* decoder is slightly modified in order to output the first 24 proposals for each submitted translation.

The *Moses* decoder is encapsulated in a Java *RMI* interface server which allows the running of several concurrent decoders (on the same or on different language pairs). Each sentence submitted is sent to the next free decoder, or queued if all decoders are busy.

This flexible configuration allows for various users to work at the same time.

The memory requirements are not necessarily an issue as our phrase tables are so big that it is impossible to store them in memory, instead they are stored on a SSD disk, which gives very good performance, in fact none of the testers has ever complained about the speed of Tapta.

The server includes a *post-processor* that deletes unnecessary spaces and recases the output taking into account the input[10].

[9] We do not want to include "out-of-patent domain" data (we experimented the inclusion of additional information like Europarl corpus but the results were rather disappointing).

[10] from the input '*DNA sequences showed by SEQ ID NOs 1 and 2*' the output of SMT is '*des séquences d' adn montrées par seq id nos 1 et 2*' which will be recased as '*Des séquences d'ADN montrées par SEQ ID NOs 1 et 2*'.

4.2 Tapta graphical user interface

The goal of Tapta is to provide a Computer Aided Translation tool, the main idea is to allow the user to drive the translation process, which is of particular importance when dealing with patent language. One of the main difficulties when translating sentences like "*A data bit to be stored in a defective cell having a logic state that is complementary to the biased logic state of the cell results in the program data being inverted and programmed*" is to find the right individual "chunks" of text for which the decoder will come up with translation proposals.

We created a Graphical User Interface developed in Java and *Swing* library. Each user has a client connected to our Tapta server, connected in turn to several *Moses* decoders (in our experiments: 3 decoders running on a multi-core server).

The user writes or copies a title and abstract into a text box, Tapta launches automatically two processes: a domain guesser and a term extraction.

The domain guesser is a program that takes all words in the input text and provides a list of related domains. This list is displayed to the user so he can select the right domain.

The term extractor parses the text and extracts adjacent words that repeat more than once. Each extracted term is displayed and pre-translated, the user can then update the translation or invalidate the term (see Figure 1).

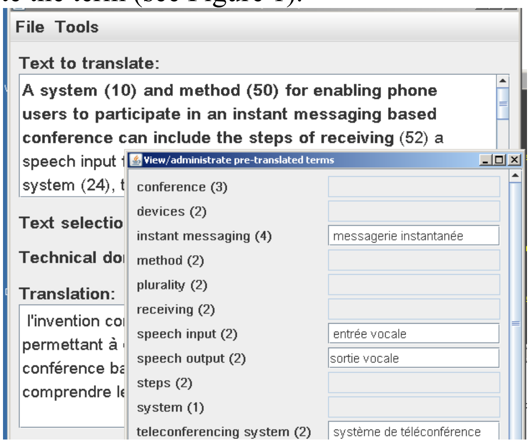

Figure 1: Tapta interface, pre-translated terms window

Once the domain and pre-translated terms are set the interactive translation can begin:

The user selects a segment of text, optionally overrides the selected domain for the segment, and submits the segment to the translation server. The server returns a ranked list of proposals (up to 24). The proposals are displayed on a menu, the user can then choose the best translation (figure 2) and the selected text is pasted on the output text box (obviously the user can edit the output text).

If the sentence is too long the user can select each individual part of the sentence to get better proposals. A "control" selection allows the user to select discontinuous terms (e.g. select '*resilient mass*' in '*A resilient liquid-impervious mass…*').

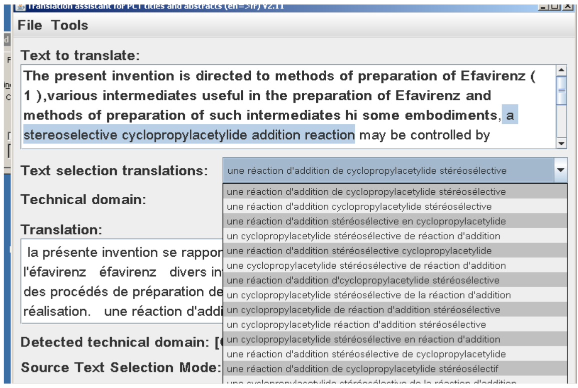

Figure 2: Tapta graphical user interface. The user selects the segment in source text box, then chooses among the proposals

Once the user has entered pre-translated terms, the source sentence is submitted using *Moses* special XML tagging to specify that the terms are already pre-translated. Each occurrence of such a term will then be uniformly translated.

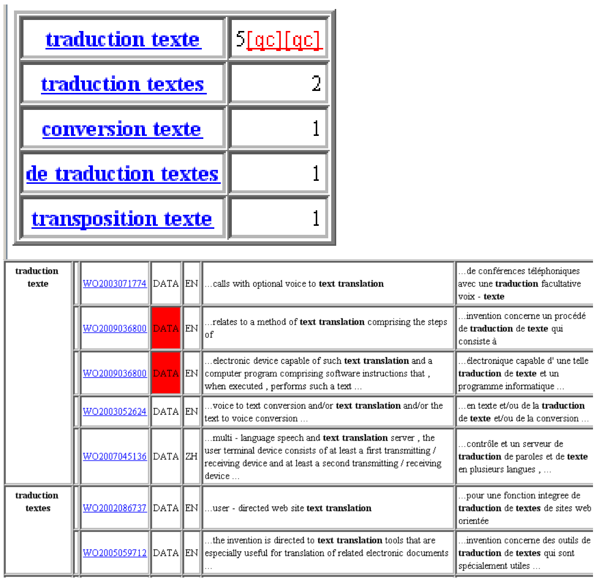

Figure 3: Concordancer for term '*text translation*', the red [qc] links to the snippet of the translation which passed the QC.

Users can access the concordancer inside Tapta. The concordancer is a Web service based on a *Lucene* index containing the information result of the word alignment (using *grow-diag-*

final-and file – *(*Koehn 2010 p. 118)). This concordancer displays the segments containing the search term and the corresponding aligned words. A first window displays the aligned words by order of frequency, so the user can immediately see which translation is the most common. The concordancer can filter translations by domain. A visual hint highlights translations that passed the QC and translations that were published initially in French (see Figure 3 for an example).

5 Results/evaluation

5.1 Automatic evaluation

The BLEU scores (Papineni 2002) were used to compare human translation with automatic proposals (with only one reference translation for each document). An attempt to optimize the settings by maximizing the BLEU score on the development set was carried out, the best settings (after a few runs with various parameters) turned out to be to use a distortion of 0.15 (in short: allows words to be easily reordered) and beam-size (s=10) reduced (in short: less proposals are kept for each word but decoding time is quicker). Table 2 gives some of our BLEU scores according to specific experiments. The parameters were first optimized on a small development set containing only 102 segments (out of 51 documents), then the settings were evaluated on a test set containing 9521 segments (out of 1390 documents), the test set better reflects the work of WIPO translation as it contains only newly published documents, new documents are likely to contain new terms, which could partly explain the BLEU difference between the test set and development set.

As expected, the QC model alone does not to produce good model (BLEU score is only 28.29).

The "QC-improved" and the domain-aware results are rather disappointing in term of BLEU improvements (a BLEU score increase of 0.01 and 0.60 respectively is not statistically significant). However we must keep in mind that Tatpa should not be judged only by BLEU on a single reference translation. A human judgment on the accuracy of proposals on a human-driven segmented text is the only good metric.

5.2 Human evaluation of Tapta driven translation

Tapta-driven translation was tested in our organisation in 2010 by the PCT French Translation Section. Two series of tests were conducted, one in March and one in September 2010. This chapter will present the tests that took place in March and their results. The September tests produced similar results.

Table 2: BLEU scores computed using various configurations and tools. ***Pruned*** **is the result of phrase table pruning,** ***BigLm*** **uses the bigger language model.** ***QC*** **means higher weight to QC-passed or French documents.** (51) **indicates the development set,** (1390) **the test set. w/oDomain means that we do not use any domain information**

Experiment	**Speed** **Seconds (#docs)**	**BLEU score**
Baseline	542 (51)	40.11
Pruned	164 (51)	41.11
Pruned+d=0.15+s=10	22 (51)	43.30
Pruned+BigLm	196 (51)	51.71
Pruned+BigLm w/oDomain	173 (51)	50.91
Only Qc	N/a (51)	28.29
Google translate[11]	N/a (51)	34.09
Bing translator[12]	N/a (51)	27.17
Pruned+BigLm	27140 (1390)	45.06
Pruned+BigLm w/oDomain	6934 (1390)	44.99
Pruned+BigLm_Qc_d=0.15 s=10	6378 (1390)	45.07

5.2.1 Set-up of the test

12 testers, with good linguistic skills in English and French but who are not professional translators nor have generally any technical background, were selected to take part in the tests. The testers were coached by two PCT translator-revisers. For the duration of the tests, the testers and the revisers were gathered in a single training room equipped with 12 PCs, one for each tester, on which Tapta had been installed. The usual terminological and linguistic resources used by PCT technical translators were also provided to the testers. Over the duration of the tests, 516 translations of English original patent documents were produced by the 12 testers.

As already stated, the testers were coached by two translator-revisers. Nonetheless, for 3 days, coaching was suspended to assess how the testers would perform without being coached. 113 translations were produced during these 3 days without coaching.

5.2.2 Results

The testers agreed that proposals were better when the domain was set properly. They also saw an improvement when the "QC-improved" model was introduced (September 2010).

[11] http://translate.google.com/ (January 2011)
[12] http://www.microsofttranslator.com (January 2011)

The 516 total translations produced by the testers were revised, partly by the two translator-revisers who did the coaching and partly by two further PCT translator-revisers. These 4 translator-revisers were instructed to assess and correct the translations done by the testers and to mark them as publishable or not publishable.

Table 3: the translations produced with and without coaching that were considered publishable (P) and non-publishable (NP) by the revisers.

	Translations produced	P		NP	
Total test period (13 days)	516	299	57.9%	217	42.1%
With coaching (10 days)	403	243	60.3%	160	39.7%
w/o coaching (3 days)	113	56	49.6%	57	50.4%

Once the translation and revision phases were completed, a few PCT experienced revisers were selected to undertake a comparative quality control (QC) of the translations produced. The purpose of this comparative QC was to compare two translations produced from the same source abstract, one by the testers, the other by outsourcing agencies currently under contract with WIPO, the comparison being kept anonymous. This comparison aimed at assessing how Tapta-assisted translations done by non-translators would be evaluated when compared to translations provided by well-established language service providers.

The comparative QC revisers were instructed to mark with P or NP the translations they had to evaluate and, when both translations of the same source text were considered to be equally P or NP ("Same evaluation" column), then to indicate which one was better ("Tapta better" and "Agency better" columns) and, if no one was better, to indicate that both were of the same quality ("Same quality" column). See Table 4 for the detailed evaluation.

The following observations and findings were deduced:

- 90.6% of the translations produced were deemed publishable after revision and, in 77.4% of the cases (50.6% + 26.8%), the translations produced were deemed, after revision, of a quality at least equivalent to that of the translations produced by the outsourcing agencies. These results were beyond expectations.
- The quality results before revision were also promising (see Table 3): 57.9% of all translations produced by the testers, a large majority under the coaching of the revisers, would have been considered publishable without revision.
- An interesting finding is that almost 50% (49.6%) of the translations produced without coaching would have been considered publishable before revision.
- It is also interesting to compare the results with and without coaching since they were not so different. Before revision, the percentage of publishable translations during the 10 days when coaching was provided was 60.3%, compared to 49.6% achieved during the 3 days without coaching.

5.2.3 Feedback

A positive feedback from the testers and the revisers alike was that since the testers were not familiar with the work and given the technical nature of the abstracts to translate, in most cases Tapta represented a useful if not indispensable aid, which allowed them to produce an acceptable translation even with no or little help from the revisers coaching them, who could not make themselves available for more than one tester at a time.

6 Conclusion and future work

Tapta proved to be an accelerated training aid, which in many instances allowed an inexperienced translator to produce rough or pre-translations of a sufficiently acceptable quality, to then be finally checked by a reviser for quality assurance or enhancement purposes. The tests

Table 4: Evaluation

Translations	translations QCed	Tapta		Agency		Same evaluation (both P or both NP)	Same quality	Tapta better	Agency better
		P	**NP**	P	**NP**				
With coaching	388	350	38	344	44	324	111	193	84
Without coaching	112	103	9	98	14	93	23	60	29
Total	500	453	47	442	58	417	134	**253**	**113**
%	**100**	**90.6**	**9.4**	**88.4**	**11.6**	**83.4**	**26.8**	**50.6**	**22.6**

also demonstrated that the time needed by beginners to prepare, draft and input the translations was reduced thanks to the many equivalents proposed by Tapta. It was also observed that the translations produced from simple source texts were often of a quality comparable to that of translations produced by translators.

Tapta has been judged as a very valuable tool for non translators: it speeds up the process of translation and also constitutes an important training aid. We have observed that the learning curve of the users of the tool is fast. For example: it is not obvious how to select the best segmentation of the input text, however after a short time, users get used to this selection process, if the sentence is simple a selection of a big segment is adequate, otherwise they select shorter segments. WIPO revisers who coached the testers agreed that, out of the 15 testers, 6 to 7 who have a more solid linguistic background were sufficiently skilled and ready to be considered as having a real and interesting potential for technical translation.

Future work includes: (a) application to other language pairs, (b) use of the *triangulation* improvement when the original document is available in more than one language (c) incremental training so that newly published translations are automatically added to the existing tool.

A first version of our Tapta tool, cut-down for the Web, has just been released on WIPO's public free patent search engine PATENTSCOPE at: http://www.wipo.int/patentscope/translate/

Acknowledgements

This work would not have been possible without the help of WIPO translators, namely Cécile Copet, Sophie Maire, Yann Wipraechtiger, Peter Smith and Nicolas Potapov. Special thanks to the 15 persons who participated in the two tests of Tapta and to Paul Halfpenny for his valuable proof-reading.

References

Federico, M., N. Bertoldi & M. Cettolo. 2008. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In proceedings of Interspeech, Brisbane, Australia, 2008

Gao, Qin and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In proceedings of the ACL'08 Software Engineering, Testing, and Quality Assurance Workshop, 2008

Johnson, Howard, Joel Martin, George Foster, Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. EMNLP-CoNLL 2007: 967-975

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris C. Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In Proceedings of ACL 07. Morristown, NJ, USA, 177-180.

Koehn, Philipp. 2009. A Web-Based Interactive Computer Aided Translation Tool. School of Informatics. University of Edinburgh

Koehn, Phillip. 2010. Statistical Machine Translation. textbook, Cambridge University Press, January 2010.

Koehn, Philipp & Hieu Hoang. 2007. Factored Translation Models. EMNLP, 2007.

Koehn, Phillip & J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In proceeding of StatMT '07

Ma, Xiaoyi. 2006. Champollion: A Robust Parallel Text Sentence Aligner. In proceedings of LREC06, (May 2006, Genoa, Italy)

Macklovitch, Elliott, Ngoc Tran Nguyen & Guy Lapalme. 2005. Tracing Translations in the making. MT Summit X, pp. 323-330, Phucket, Thailand, oct 2005

Macklovitch, Elliott. 2006. TransType2: The last word. In proceedings of LREC06, (May 2006, Genoa, Italy)

McCandless, Michael, Erik Hatcher, Otis Gospodnetić. 2010. Lucene in Action. 2nd Edition. Manning Press.

Niehues, Jan and Alex Waibel. 2010. Domain adaptation in statistical machine translation using factored translation models. In proceedings of EAMT 2010, 27-28 May 2010, Saint-Raphaël, France

Papineni, K., S. Roukos, T. Ward, and WJ Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In proceedings of ACL 2002, pp. 311-318

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

Thurmair, Gregor. 2009. Comparing different architectures of hybrid machine translation systems. 2009. In proceedings of the 12th Machine Translation Summit, August 26-30 2009, Ottawa, Canada; pp.340-347.

Tsai, Yvonne. 2008. Supply and Demand Analysis of Patent Translation, Translation Journal Volume 12, No. 3 July 2008

WIPO. 2010. PCT The International Patent System - Yearly review, developments and performance in 2009, WIPO Publication No. 901(E)/09, June 2010