Automatic translation tools at WIPO

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

Introduction

The World Intellectual Property Organization has access to millions of patent applications written in several languages (the abstract of each "PCT application" has to be available in at least French and English). With our search engine PATENTSCOPE, users can search and browse this collection in addition to other national collections in various languages (English, French, Spanish, German, Chinese, Japanese, Korean and Portuguese).

A statistical machine translation (SMT) software tool can use our huge parallel corpora to learn different language pair translation models, as carried out here at WIPO, in addition to a "cross-lingual information access" tool, WIPO now offers users an assistant to translate patent applications and a tool more specifically targeted at assisting translators to accelerate the translation of patent applications in an interactive way.

Background

Statistical machine translation (Koehn 2010) is increasingly applied in the field of patent translation: Machine Translation Task at NTCIR-9[1], the EU-funded project Pluto (Tinsley et al. 2010) and the collaboration between the European Patent Office and Google translate[2] (Täger 2011), etc.

WIPO has experimented with the use of open-source software Moses (Koehn et al. 2007) and now provides three tools:

1) cross-lingual search for users: CLIR
2) computer aided translation tool for translators: TAPTA-js, and its Web version TAPTA-Web
3) web-based gist-translator (targeting PATENTSCOPE users): TAPTA-Web-Lite

These three tools make use of the data-driven approach and use the classification of the applications (International Patent Classification - IPC) to take into account the domain when accessing translation proposals (the term 'automatic translation' will be generally translated into French as 'traduction automatique' but will be translated as 'translation automatique' in a domain like mechanical engineering).

PATENTSCOPE and cross-lingual search: CLIR

PATENTSCOPE is WIPO's patent application search engine. As it contains patent applications in various languages, we offer users the possibility to search in multiple languages.

A SMT model for various language pairs was trained using mainly patent application titles, we then automatically extract bilingual terminology from this model. CLIR allows users to search a term or a phrase and its variants in English, French, German, Japanese, Spanish, Chinese, and Korean (and to some extend Russian and Portuguese too) just by entering the term(s) in one of those languages in the search box. The system will suggest variants and translate the term(s), allowing the user to search

[1] This task proposes that participants 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
[2] See http://www.epo.org/topics/news/2010/20101130.html

patent documents which were disclosed in a foreign language. CLIR can also be used to look for synonyms in different domains.

Querying "toothbrush" in PATENTSCOPE in English language with the following query:

```
EN_ALLTXT:("toothbrush")
```

PATENTSCOPE queries are boolean queries combining field name(s) (here EN_ALLTXT means "all texts in English language")

=> This query returns 7433 documents (7/10/2011)

While using the "CLIR" search engine the query becomes:

```
EN_ALLTXT:("toothbrush" OR "tooth brush")  OR
DE_ALLTXT:("Zahnbürste") OR
ES_ALLTXT:("cepillo de dientes" OR "cepillo dental") OR
FR_ALLTXT:("brosse à dents") OR JA_ALLTXT:("歯ブラシ") OR
KO_ALLTXT:("칫솔") OR PT_ALLTXT:("escova de dentes") OR
RU_ALLTXT:("зубная щетка") OR ZH_ALLTXT:("牙刷")
```

This query combines various languages containing translation(s) (e.g. 'cepillo de dientes' and 'cepillo dental' in Spanish) and also synonyms (e.g. 'toothbrush' and 'tooth brush' in English)

=> This query now returns 11,597 documents (7/10/2011)

This tool is available as part of the general PATENTSCOPE search engine at http://www.wipo.int/patentscope/search/clir/clir.jsp

Computer aided translation tools: TAPTA suite

A SMT was trained using titles and abstracts (aligned at sentence or segment level). This SMT model is of very high quality thanks to the amount of training data (especially true for English-French: with more than 8 Million parallel translation units as released in COPPA corpus[3]). We called these tools TAPTA (Translation Assistant for Patent Titles and Abstracts). We developed three versions targeting different users: TAPTA-js (targeting WIPO's internal users), TAPTA-Web (targeting external translators) and TAPTA-Web-Lite (targeting general users).

WIPO's internal interactive translation: TAPTA-js

An interactive graphical user interface was created that allows users to drive the translation interactively (selecting the best segments to translate and choosing the right proposal). The first version was developed as a Java-swing application (hence the name). A significant experiment was conducted with human operators and the tool has been used to help non-professional translators to translate patent applications. The output was judged to be successful. All details have been published in the paper "Tapta: a user-driven translation system for patent documents based on domain-aware statistical machine translation" (Pouliquen et al. 2011).

[3] WIPO has recently released a product containing all aligned titles and abstracts translated in the PCT (Pouliquen & Mazenc 2011). This corpus contains more than 8 Million translation units and can be used to feed translation memory or to train a statistical machine translation tool (the corpus is free for research). Address: http://www.wipo.int/patentscope/en/data/products.html#coppa

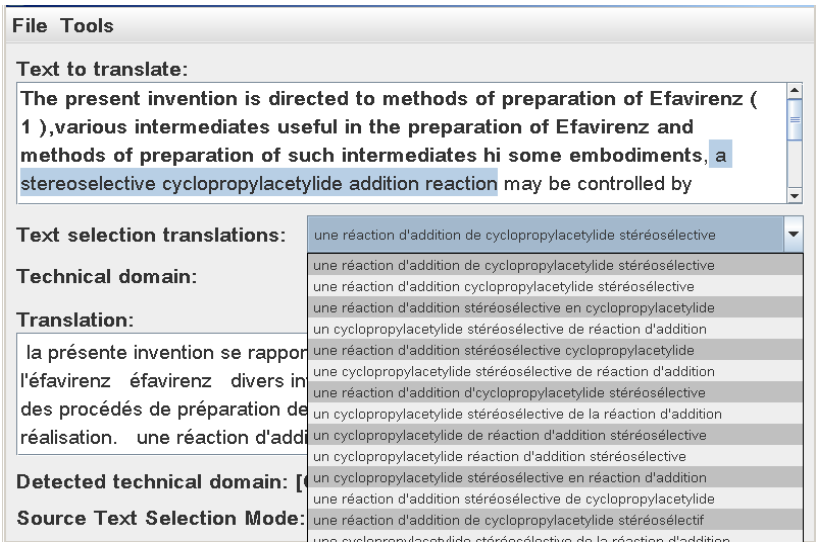

Figure 1: TAPTA-js (Java Swing) interface: the user highlights the next segment he wants to translate and gets translation proposals

WIPO's external interactive translation: TAPTA-Web

A version of Tapta has been adapted to the Web, with the same interactive mode. A translator can then drive the translation himself. Basically the user selects the next segment to be translated (with the mouse or keyboard) and can then choose alternatives among the proposals.

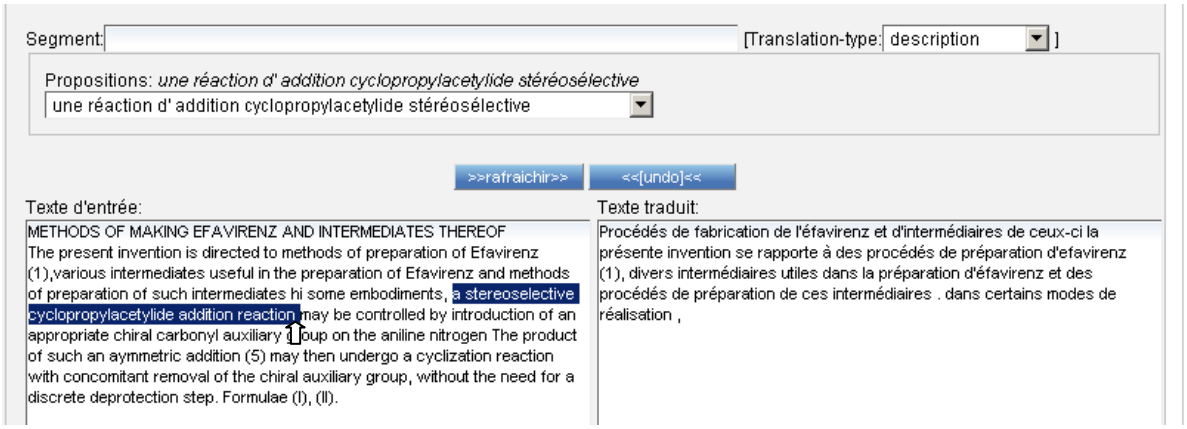

Figure 2: TAPTA-Web interactive

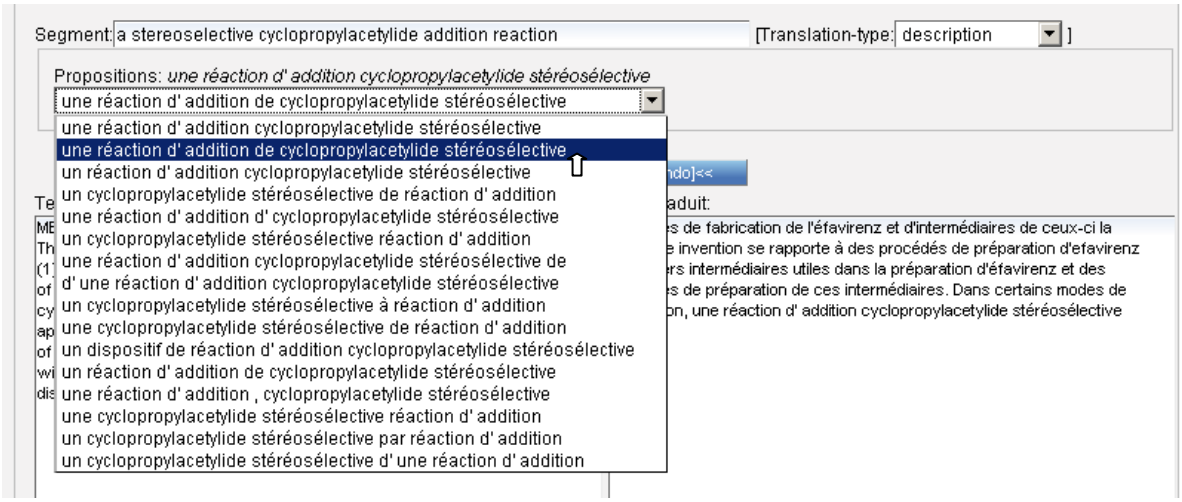

Figure 3: TAPTA-Web, session example

A demonstration of the tool will be made live during the presentation. Figure 2 and Figure 3 show a translation process example

WIPO's external gist-translator: TAPTA-Web-Lite

In PATENTSCOPE, the user may retrieve a patent application which is only available in languages he does not understand very well. The PATENTSCOPE website offers users the possibility to automatically translate using freely available widgets: Google translate or Microsoft translator[4] (see Figure 4 for examples of translations).

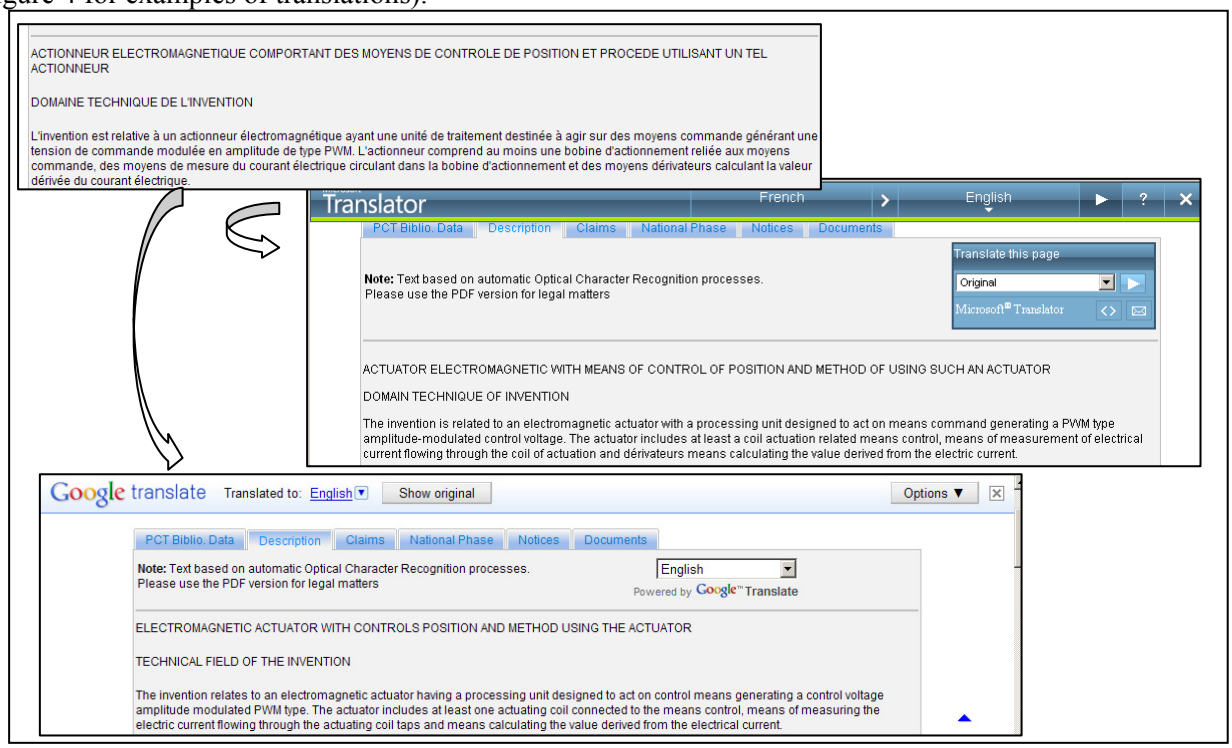

Figure 4: Displaying Google/Microsoft translations

As described in the two previous sections, WIPO has its own machine translation tool, we decided to make it available for general users as an alternative gist-translator.

[4] Option recently added (October 2011)

This new tool, called TAPTA-Web-Lite, aims at offering users a gist-translation of a patent application (currently only title and abstract). It can translate texts from English to Chinese and French and vice-versa (more languages to be added in the future). This tool allows the user to browse the segments of the translated text and access other proposals (the user can also provide his own segmentation of the source text and type in his own translations).

The TAPTA-Web-Lite tool is publically available at http://www.wipo.int/patentscope/translate, any user can test it with title and abstracts taken from Patent applications. The user can access various proposals for one translated segment (see Figure 5) and he can isolate a specific segment (see Figure 6) to get even more proposals.

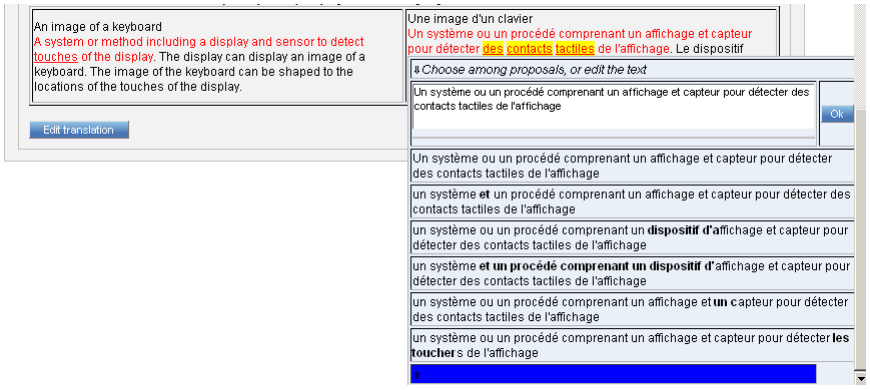

Figure 5: TAPTA Web-Lite, listing translation proposals

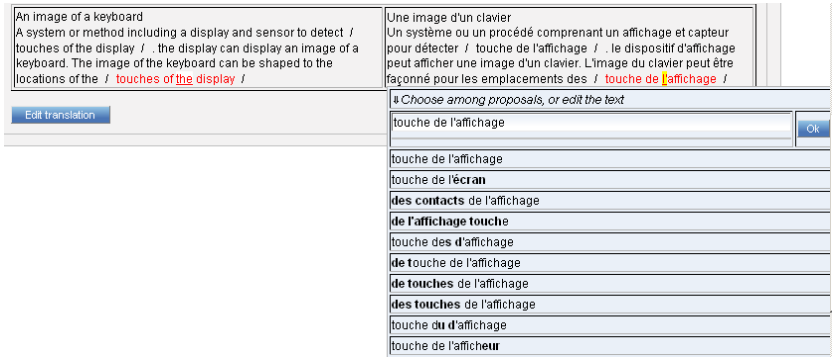

Figure 6: TAPTA-Web-Lite, isolating a specific phrase in order to get more proposals

From 15 September to 15 October 2011: we received, on average, 400 translation requests per day, 58% of the requests come from Chinese IP addresses, Chinese to English being the most used direction (34%) followed by English to French and English to Chinese (both 29%) followed by French to English (8%).

We expect to have more requests as we recently fully integrated TAPTA-Web-Lite as part of PATENTSCOPE's search engine. The user can now access the French or Chinese automatic translation when he chooses French or Chinese as interface language and the abstract is not available in his language, see Figure 7.

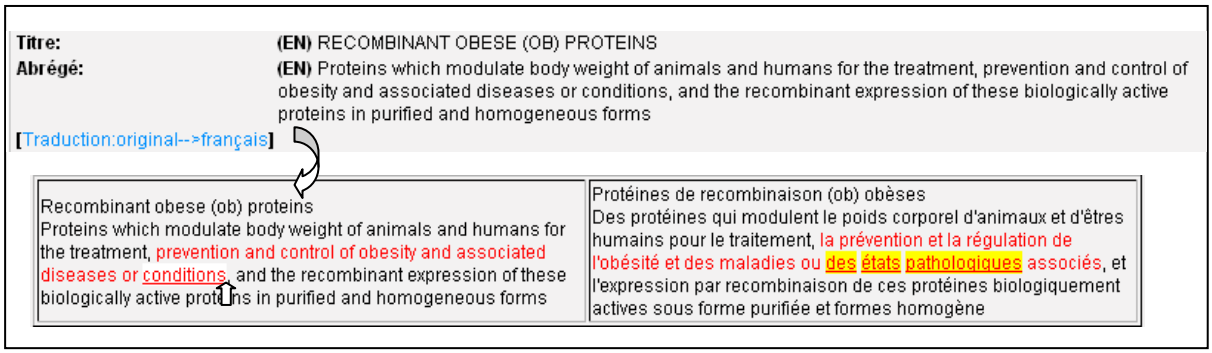

Figure 7: Accessing TAPTA-Web-Lite translation from PATENTSCOPE (the user has selected the French interface and the abstract is not available in this language)

Conclusion

Exploiting parallel corpora (like the one we use) can create useful tools for both translators and users. The underlying models are built completely automatically, which allow us to add new language pairs as soon as parallel data becomes available.

The translation tool suite TAPTA is still at prototype level, however we have already published the TAPTA-Web-Lite tool on the Web, this version is now used daily (400 requests per day) and we receive positive feedback from users.

We want to continue to work on:

1) improving our tools
2) integrating external translation tools (Google translate, Microsoft Translator, may be in the near future a Korean translator: KIPO translate)
3) pushing forward innovation in the domain of patent translation (for example, the release of COPPA corpus)

Bibliography

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

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.

Pouliquen, Bruno, Christophe Mazenc (2011) COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent language barrier at WIPO [MT summit 2011] Proceedings of the 13th Machine Translation Summit, p 24-30, September 19-23, 2011, Xiamen, China.

Pouliquen, Bruno, Christophe Mazenc & Aldo Iorio, (2011) Tapta: a user-driven translation system for patent documents based on domain-aware statistical machine translation. *[EAMT 2011]: proceedings of the 15th conference of the European Association for Machine Translation*, 30-31 May 2011, Leuven, Belgium; eds. Mikel L.Forcada, Heidi Depraetere, Vincent Vandeghinste; pp.5-12.

Täger, Wolfgang, (2010): The sentence-aligned European patent corpus. *[EAMT 2011]: proceedings of the 15th conference of the European Association for Machine Translation*, 30-31 May 2011, Leuven, Belgium; eds. Mikel L.Forcada, Heidi Depraetere, Vincent Vandeghinste; pp.177-184.

Tinsley, John, Andy Way and Páraic Sheridan, (2010) PLuTO: MT for online patent translation. [*AMTA 2010]: the Ninth conference of the Association for Machine Translation in the Americas*, Denver, Colorado, October 31 – November 4, 2010; 8pp

WIPO. (2010). PCT The International Patent System - Yearly review, developments and performance in 2009, WIPO Publication No. 901(E)/09, June 2010 available at http://www.wipo.int/pct/en/activity/pct_2010.pdf