Full-text Patent translation at WIPO: scalability, quality and usability

Bruno Pouliquen Bruno.Pouliquen@wipo.int
World Intellectual Property Organization, Global Databases Service
34, chemin des Colombettes, CH-1211 Geneva 20, Switzerland

Abstract

WIPO has access to a huge amount of patent application texts in different languages, therefore, we have built a machine translation tool (called WIPO translate) trained on big parallel data. We focus on offering quality machine translation to the general public, WIPO translate is fully integrated on our search engine PATENTSCOPE[1]. We have recently experimented aligning patent full-texts (description and claims) and our tool can now be trained on billions of words. Automatic evaluation metrics show an improvement over publicly available translation sites for the translation of patent texts (e.g. Google Translate, Microsoft translate). We have developed specific user interfaces, which are fully integrated, in our search engine PATENTSCOPE with reasonable translation speed. WIPO translate has now reached maturity in providing translation with competitive scalability, quality and usability.

1. Introduction

WIPO has 5 years' experience in providing quality machine translation on its search engine PATENTSCOPE. Originally trained exclusively on patent titles and abstracts, we have now experimented using descriptions and claims (full text) to train our statistical machine translation tool (called WIPO translate), based on the open source toolkit Moses. Machine translation of patent texts is now integrated in PATENTSCOPE, despite the issues of scalability (translation models trained on billions of words), quality (our automatic evaluation shows an improvement over publicly available translation sites: e.g. Google Translate) and usability (it is fully integrated in our search engine PATENTSCOPE, with a translation speed of less than 2 seconds per sentence).

2. Background

The World Intellectual Property Organization (WIPO) provides access to about 50 million patent applications on its search engine PATENTSCOPE. It includes the international Patent Cooperation Treaty (PCT[2]) applications, but also documents of participating national and regional patent offices.

All the PCT applications must be filed in one of the following languages: Arabic, German, English, Spanish, French, Russian, Japanese, Korean, Portuguese or Chinese, then the title and abstract are translated into English and/or French. The description and the claims remain available only in the original filling language.

[1] http://patentscope.wipo.int
[2] http://www.wipo.int/pct

The other applications (from regional or national patent offices) are usually published only in the original language (sometimes the title and the abstract are translated in one other language, rarely the claims and almost never the description).

This creates a huge need for users to read patent applications in a language they do not master. WIPO has investigated the use of machine translation to offer users a tool to help them understanding the content of any patent application.

WIPO previously investigated the use of machine translation to offer users the possibility to translate the title and abstract of patent applications (see section 2.1). However, this first solution was not suitable for the description and claims.

As a temporary solution, PATENTSCOPE offered the possibility to translate description and claims using public translation engines: Google translate, Microsoft/Bing translate and, recently, Baidu translate.

A recent study shows that, every day, about 5 million words are automatically translated using one of the machine translation tools available on PATENTSCOPE (Google translate/Microsoft translate). It clearly indicates the need for such tools in the intellectual property domain. This does not come as a real surprise as, of the total of about 32 million descriptions available on PATENTSCOPE, (as of September 2015), only 11.5 million (36%) are written in English (see Table 1)

Language	***# applications (millions)***	***Percentage***
English	11.5	36.16%
Japanese	8.2	25.79%
Chinese	4.5	14.15%
German	2.8	8.81%
Korean	2.7	8.49%
Spanish	1.1	3.46%
Russian	0.7	2.20%
French	0.3	0.94%
Portuguese	0.2	0.63%
total	32.0	

Table 1: Number of patent applications in PATENTSCOPE having their description available in a specific language

2.1. Statistical machine translation

The statistical machine translation (SMT) approach "learns" its translation model using parallel sentences and then combines it with the target language model learned on monolingual texts. This fully automatic approach is suitable for the patent domain as we can automatically build such parallel corpora.

WIPO translate is based on the open source Moses[3] (Koehn et al 2006). We have built a set of tools to pre-process the texts and to offer practical interfaces to Moses. WIPO tools include specific natural language processes: decompounder (German, Korean), pre-reordering (German, Japanese), prefix splitting (Arabic), Tokenizer (Chinese, Japanese)…

[3] www.statmt.org/**moses**

The first version of WIPO Translate (Pouliquen et al. 2011) was trained on 180 million words (8.7 million English and French segments). The corpus has been released for free for research purposes ("COPPA corpus" Pouliquen & Mazenc, 2011[4]).

2.2. Domain aware machine translation

We make use of the IPC classification[5] to categorize any application into 32 domains (medicine, data, engineering, chemical …); this domain information is then encoded as a "factor" in Moses so that the phrase table can give "priority" to in-domain phrases. WIPO translate can then translate differently the same sentence in any of the 32 domains. Similarly, we decided to include the fact that a sentence belongs to the description or to the claim as a factor in our phrase table. Various experiments are still going on to better include this context information in the translation process.

3. Description of the tools

3.1. Text alignment

We need texts aligned at the sentence level in order to train Moses models. This is straightforward for titles when an application has a title in two languages. We need to apply better techniques for abstracts (e.g.: one sentence in English could be translated as two in French); the WIPO home-made sentence aligner has been developed for this purpose.

For descriptions and claims, the problem is different: the same invention can be submitted in different offices in different languages, but the description may be slightly different and the claims are often re-written according to the office and the protection needed by the applicant (see for example Täger 2011). Links between applications of the same invention are stored in a "priority list", but the parallel corpus one can extract from this information is rather a "comparable" corpus than a real parallel corpus. WIPO adapted its sentence aligner tool to better filter descriptions and claims. WIPO aligner relies on bilingual dictionaries to automatically align sentences of noisy comparable corpus. It can discard non-aligned set of sentences or discard noisy texts (where the texts are not any more a translation of each other). Similarly, we heavily filter the claims and do not try to align them when the number of claims is different (e.g.: if a Chinese Patent Office application contains 11 claims then the corresponding US Patent Office application must also contain 11 claims) or when the intra-claims references are different (e.g.: if the third Chinese claim refers to the first claim then the third US claim must also refer to the first claim).

3.2. Technical infrastructure

The web application is distributed via a software load balancer to two servers. Each server calls a set of "WIPO translate servers" which in turn call a set of "Moses Engine servers". This architecture allows for a robust and scalable set up where we can build Moses translation server farms (adding translation servers when new language pairs – or more engines for an existing pair - need to be added). It also avoids any single point of failure: if a web application or a translate server fails, the application can continue working.

Confidentiality is of high important for PATENTSCOPE users (a private company may not want its translation requests to be observed by another company, e.g. Google or Mi-

[4] The COPPA corpus is available at: http://www.wipo.int/patentscope/en/data/#coppa. Note that the Version 2 (to be released in 2016) will include more applications and more languages.
[5] http://www.wipo.int/classifications/ipc

crosoft) therefore all our servers work in https mode. This ensures users that no information (including IP address) is ever disclosed.

It should be noted that WIPO translate includes automatic monitoring and alerting tools to ensure a good customer service close to 24/7 (with minimal administration work).

3.3. Handling large data: scalability

The translation models, even when they are trained on big data, must be of a "reasonable" size. WIPO uses a set of filters, pruning processes and binarization tools in order to keep the model size to a minimum, without sacrificing much quality. See Table 2 for an example of size reduction: pruning and binarization manages to reduce a 342 Gb model to 15.2 Gb (4.4% of the original size).

	Phrase table		**Reordering model**		**Language model (5 grams)**		**Total size**
	# rows (in million)	**Size**	**# rows** (in million)	**Size**	**#ngrams** (in million)	**Gb**	
Basic	806	100.0G	806	89.0G	584	23.0G	342.0Gb
Pruned	551	69.0G	551	61.0G	388	16.0G	
Binarized		6.4G		4.2G		4.6G	15.2Gb

Table 2: size reduction (Chinese into English model)

WIPO translate must offer translation of any text in a "reasonable" time. We are trying to parallelize the decoding process and avoid time-consuming analyzers. We can now translate a full page within few seconds.

3.4. Quality

We use various automatic metrics (BLEU, METEOR, RIBES) to compare different versions of WIPO Translate, but also to compare WIPO translate to other engines output. We conducted an evaluation of the translation of patent application texts (1000 randomly selected sentences from newly published patent applications) the same text was submitted to WIPO translate and Google translate (see Table 2 for the results). Note that we use only title and abstracts, but, for Chinese, we conducted an evaluation on claims and descriptions.

From language into English	***WIPO translate***	***Google translate***
German title&abstract	46.11	37.94
Spanish title&abstract	36.00	33.07
French title&abstract	46.97	41.72
Russian title&abstract	28.88	17.76
Korean title&abstract	22.09	19.85
Japanese title&abstract	22.10	21.27
Chinese title&abstract	26.37	21.80
Chinese claims	28.68	21.89
Chinese descriptions	38.03	32.40

Table 3: Comparison between WIPO translate and other engines (BLEU scores)

3.5. Usability

WIPO translate aims at offering users access to automatic translation of patent texts, therefore we try to give easy access to translation tools to PATENTSCOPE users.

3.5.1 Cross lingual Information Retrieval (CLIR)

The CLIR allows users to search a term or a phrase and its variants in English, French, German, Spanish, Portuguese, Japanese, Russian, Chinese, Korean, Italian, Swedish or Dutch by entering the term/s in one of those languages in the search box. The system will suggest variants and translate the term(s), therefore allowing users to search PATENTSCOPE for documents disclosed in a language that they do not master. This system has also been automatically trained using Moses on titles and abstracts.

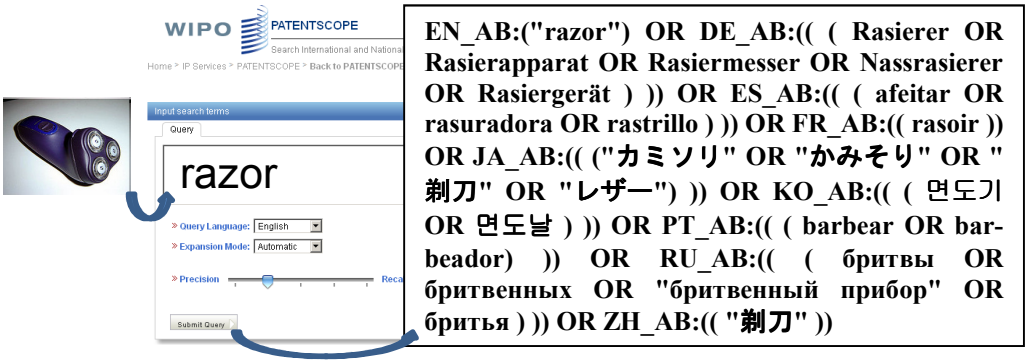

Figure 1: Example of CLIR query ("razor" automatically expanded to various languages)

3.5.2 Translating short texts: web interface

Any user can access WIPO translate from a simple web interface, publically available at https://patentscope.wipo.int/translate, this allows for the translation of any given text.

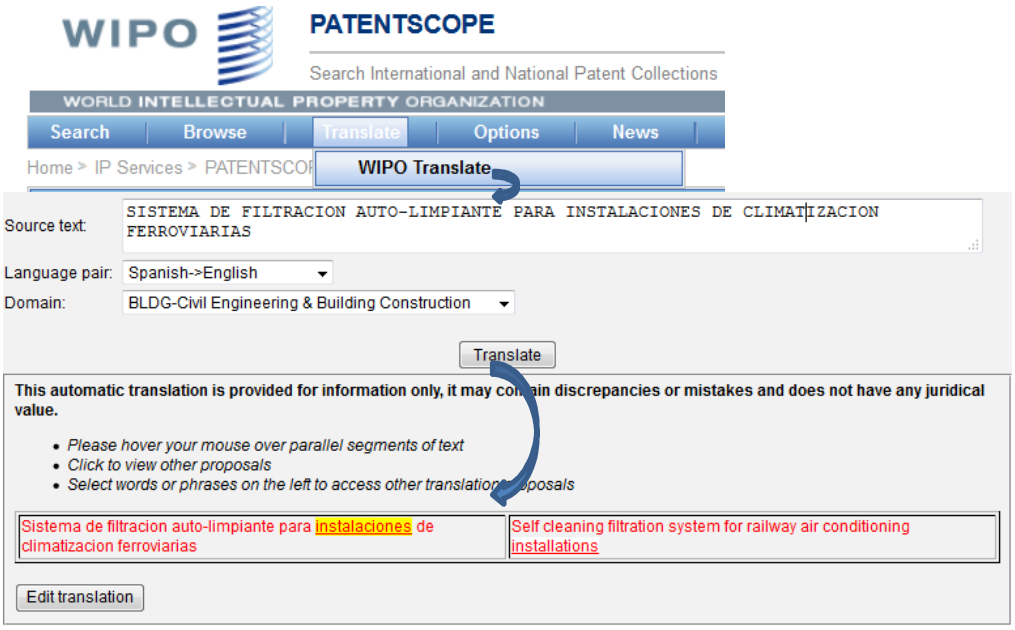

Figure 2: WIPO translate web interface to translate short texts

3.5.3 Translating long documents: WIPO translate widget

Translation of patent application descriptions created a new challenge, as the texts are usually long (on average 6,400 words, but sometimes up to 150,000 words). Launching the translation of the full document would require many CPUs for each language pair.

It should also be noted that pre-translating all the texts is not really an option as our models are evolving over time and PATENTSCOPE has an enormous corpus of texts[6] that would require weeks to be translated.

Therefore WIPO has developed its own "widget" (JQuery program running on the client side) which is used to translate only the sentences the user is currently reading on the screen. This technique allows for a better share of the server load among users of PATENTSCOPE. It has been decided, as a first step, to offer only translation of Chinese-English (both directions), but more languages will be added in the future.

[6] The total number of characters of English text in PATENTSCOPE is about 1 trillion (to give an order of idea, English Wikipedia is 49 billion characters)

Figure 3: WIPO translate widget example, online translation of a Chinese description

This widget can be smoothly included on any page, it is currently available on the description page, but also for the claims page, the bibliographic page and the result list.

3.6. Our software in other contexts

It is of note that the software has been successfully installed in other UN organizations. It has been trained on their own data and is now running with different models, different usage scenarios and different quality. In contrast to WIPO Translate on PATENTSCOPE (where the tool is used for assimilation: i.e. to offer users a "gist" translation) the tool in other contexts is used mainly for disclosure, that is to say as a "translation accelerator", offering translators a first translation to refine.

WIPO translate software (called "TAPTA") can handle the internal parallel documents of an organization and can then offer translators a quality machine translation reproducing their own internal jargon. It has been installed at the United Nations since 2011 (see details in Pouliquen et al., 2013) and recently at the International Maritime Organization (IMO) (see Pouliquen et al. 2015). The tool is used internally in WIPO PCT (for the translation of titles and abstracts), in the Madrid sector (to translate goods and services) and in the WIPO Language Division (to translate official documents). In addition, it has been installed at the Food and Agriculture Organization (FAO) and the International Telecommunication Union (ITU). Early prototypes have been installed at the World Trade Organization (WTO), the International Labour Organization (ILO) and the International Social Security Association (ISSA).

4. Conclusion and future work

WIPO translate has now been running daily for five years, and we have smoothly incorporated the "WIPO translate widget" and the translation of description and claims. Despite the challenges of scalability, quality and usability, WIPO translate has reached maturity and is now a reliable system.

The WIPO translate widget has been available online since the beginning of September 2015, we can already see a major increase in the number of words translated every day[7], even if we currently offer only Chinese-English translation for descriptions and claims.

We hope to extend full-text document translation to new languages in the next future (Japanese, German etc.) provided that we get enough computer power. A preliminary evaluation shows that we can get big models using Japanese (estimation: 3 Billion words), German (~ 1 Billion words) or Russian (~ 200 Million words). We also plan to offer Arabic and Portuguese translation in the future (our first goal is to cover the 10 official languages of the PCT).

We are currently investigating the following topics:

- Translating through pivot language to offer any language combination (e.g. German-Japanese)
- Various techniques for a better domain adaptation (e.g. use different language models for descriptions/claims/IPC domains etc.).
- Use collected post editions to add quality estimation metric for each translated sentence
- Use transliteration techniques to "translate" applicant names across different scripts: Arabic, Latin, Cyrillic, Hangul (Korean alphabet), Chinese, Kanji+Hiragana+Katakana (Japanese) etc…

Acknowledgement

Many thanks to Marcin Junczys-Dowmunt for his indispensable contributions to this work, and to Christophe Mazenc who built the CLIR component and for managing this project.

References

Koehn P, Hoang H, Birch A, C. Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation, proc of ACL 2007. Morristown, NJ, USA, pp. 177-180.

Pouliquen, B, Mazenc C & Iorio A (2011), Tapta: a User-Driven Translation System for Patent Documents based on Domain-Aware Statistical Machine Translation, *proceedings of EAMT 2011* , Leuven, Belgium; 30-31 May 2011, pp. 5-12

Pouliquen B & Mazenc (2011) C: COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent barrier at WIPO. *MT Summit XIII: the Thirteenth Machine Translation Summit Asia-Pacific Association for Machine Translation (AAMT)*, 19-23 September 2011, Xiamen, China; pp.24-30

Pouliquen B, Elizalde C, Junczys-Dowmunt M, Mazenc C & Garcia-Verdugo J (2013): Large-scale multiple language translation accelerator at the United Nations. *Proceedings of the XIV Machine Translation Summit*, Nice September 2-6, pp. 345-352

Pouliquen B, Junczys-Dowmunt M, Pinero B, Ziemski M (2015) SMT at the International Maritime Organization: Experiences with Combining In-house Corpora with Out-of-domain Corpora. *Proceedings of EAMT 2015 conference*, Antalya, Turkey, May 2015

Täger W (2011), The Sentence-Aligned European Patent Corpus, *proceedings of EAMT 2011*, Leuven, Belgium; 30-31 May 2011

[7] In September 2015, WIPO translate receives an average of 200,000 words/day while, before the introduction of the translation of description and claims, the rate was 50,000 words/day. The translation of descriptions (English from/to Chinese) accounts for half of the requests.