Bootstraping a Comparable Corpus from Patent Family Members

Mihai Lupu
ESTeam AB / Vienna University of Technology
Sweden / Austria
mihai@mihailupu.net

***Abstract*—We present a method to generate comparable corpora from different patent documents covering the same invention. We rely on the fact that many inventors apply for protection in more than one jurisdictions. Often, these jurisdictions have different publication languages, and therefore, the same invention is described in more than one language. We use this fact to generate comparable corpora in any language pair where patent documents are available. We do this at the level of the title, abstract, description and claims and present statistics for English-Spanish data thus generated. We then show that with an additional filtering step we can reduce the errors inserted in the collection by the automated procedure.**

I. Introduction

The patent system is fundamentally a global one. Novelty, the main criteria for patentability, is language independent. At the same time, with very few laudable exceptions, patents are issued only in the language of the patenting office, and searches are often restricted to the language of the searcher, or, at best, a small subset of languages. Statistical machine translation systems are the only way forward in coping with this large multilingual corpus. The problem is that the nature of the text in the patents (both scientific and legal at the same time) [1] makes an off-the-shelf translation system perform at a much reduced rate than on a language-wide representative corpus [2].

To obtain training material for SMT systems, we cannot rely on a parallel corpus but for a very small set of languages. A comparable corpus however is almost readily available. It is the result of manual work performed by trained multilingual professionals, when an invention aims for protection in different countries. These manual translations are a highly expensive process, whose results are, in principle, publicly accessible. The results of this process are buried in different patents and patent offices, difficult to reach by the public. The availability of patent data in digital form makes this process significantly easier, but not obvious. The work described in this paper shows how to automatically obtain a comparable corpus from a large patent database. In this sense, we structured the presentation as follows: Section I-A provides an overview of the patenting system, with focus on those aspects used in this work. Section II describes the data collection, as well as the general method used in extracting comparable paragraphs. We then show a use-case for Spanish-English data in Section III. Related work and Conclusions are in Sections IV and V respectively.

A. The Patent System

To facilitate understanding the characteristics of patent corpora, we need to establish the terminology used in the patent domain.

A patent is a set of exclusive legal rights, for a limited period of time, for the use and exploitation of an invention in exchange for its public disclosure. A common first step in the patenting process is to file a patent application with a patent office. The applicant must supply a written specification of the invention (i.e. a *patent application document*) where the background of the invention, a description of the invention, and a set of claims defining the scope of protection, are given. The patent application is examined by the office where it was submitted and a patent is granted if all conditions are met.

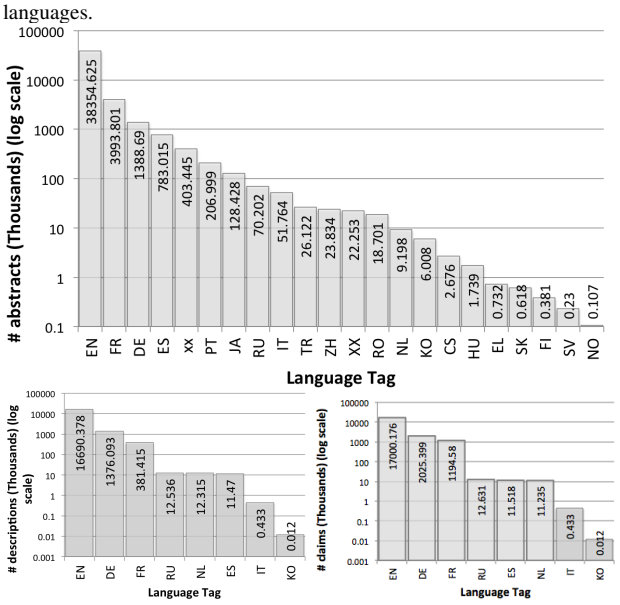
Patent documents generated at the different stages of the patent's life-cycle are identified by a *country code* (denoting the patent office), a *numeric identifier*, and a *kind code* together with a version number. Together, these three components form a unique *global* identifier.

To protect an invention in several geographical areas, a patent application can be filed at more than one patent office. The legal system allows an applicant to claim priority on a particular invention, if a patent for it has been applied for at any other patent office around the world, within a time frame (generally 18 months). As most large companies apply for protection in several jurisdictions, many patents have in their metadata the 'priority number', which is the identifier of the first patent applied for that invention. When the same invention is granted a patent by different patent offices, the two patents are said to belong to the same *patent family*. What exactly is a patent family is subject to interpretation[1], but they all are a function of the priority numbers assigned to each patent. The use of patent families provides a more extensive set of languages. The caveat is that, in situations where a patent application to one patent office is split into several at another, or vice-versa, it is no longer clear which parts match which other parts at the other patent offices.

Only one office issues granted documents in more than one language: the European Patent Office requires all applicants, upon having granted them a patent, to provide

[1] http://www.epo.org/searching/essentials/patent-families/definitions.html

Figure 1. Numbers of abstracts, descriptions and claims in different languages.



translations of their claims in English, French and German. The rest of this vast multilingual data set is hidden in different documents linked only by priority numbers.

II. Data and Methods

The method proposed here relies on the existence of digital patent information. While such data is generally available for recent patents from individual patent offices, obtaining it from each and every office is difficult. In this case, we have used the Alexandria Patent Data Warehouse [3], kindly made available to us by Fairview Research, and currently available as IFI Claims®Global Patent Database[2].

A. Data

In the version we used[3], a copy made in early 2011, the Alexandria Patent Data Warehouse covers over 70 patenting authorities for a total of just over 72 million patent documents.

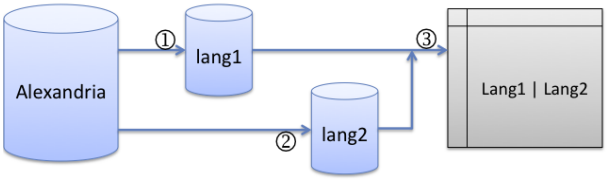
Each of the sections of a patent contains a language tag identifying the language of that particular instance. As one patent document may contain the same section in different languages, this is stored within the XML node of every section. These language tags allow us to quickly identify the useful materials within the collection. Figure 1.[4] shows, for each of the three sections, the number of documents per language. Table I. shows the number of documents having a section in at least two languages. Note that this table does not contain a *descriptions* section, because there

[2]http://www.ificlaims.com

[3]DISCLAIMER: the observations presented here refer to a specific version of the Alexandria data, are of the author's only and do not necessarily reflect the contents of the current repository, nor those of any other subsets or versions.

[4]The *xx* language tag in Figure 1. denotes missing data

Figure 2. Phase 1: matching sections together



are no documents with parallel descriptions. All values are presented in thousands, and Table I. shows only those pairs which are supported by at least ten thousand documents.

Comparing Figure 1. with Table I. we can see that relying only on the existing parallel corpus, is missing out on a large set of multilingual data. In what follows, we present a method to link different documents together via their patent family identifier, and create a collection of comparable texts, for any two languages in the collection.

B. Method

The idea is simple: using the family identifiers, select those patent sections which belong to the same family and have different language tags. The actual process involves two phases, each of at least three steps. Phase 1 creates a table with pairs at section level (e.g. two full descriptions on each row). Phase 2 splits this into paragraphs and matches the paragraphs in the two languages chosen.

1) Section-level matching: Matching sections from the same family involves a self join on the main patent table, which has just over 72 million entries, and a join with the corresponding section table. In the case of the abstracts, this has approximately 90 million entries. To make this process more efficient, for each language pair of interest (*lang1*,*lang2*), we extract the section pairs in three steps, as depicted in Figure 2.:

(**1**) extract all sections with the language tag corresponding to *lang1* and populate the table with family identifiers.
(**2**) repeat step 1 for *lang2*.
(**3**) join the two tables thus created on the family identifier.

The result of this phase is in practice of limited utility. Except for abstracts, the other sections are too large, and too prone to change across different patenting authorities, to represent a reliable corpora. The abstracts are small enough,

	Abstracts			Claims		
	EN	FR	DE	EN	FR	DE
FR	3584	-	436	886	-	886
DE	994	436	-	886	886	-
ES	165	14	0	0	0	0
JA	121	121	0	0	0	0
RU	35	1	0	0	0	0
IT	10	0	0	0	0	0

Table I
Parallel sections within the corpus (Thousands)

and of relatively little importance, that they often remain the same after translation [4]. Claims on the other hand are changed as a function of the granting practices of each office. Similarly, descriptions may change in time, function of developments in the technical field. This is why we go more in depth and look at the paragraphs within.

2) Paragraph-level matching: The problem with having pairs only at section level is that in the case of the descriptions and claims, these may be extremely long [5]. Phase 2 of our method involves 4 steps, as follows:

(**1**) **start** by identifying paragraphs based on the XML tags present in the text and matching them based only on their size. Here, we consider a pair to be a candidate simply if their size difference is less than 20% of each of the two paragraphs. We compute size as the number of characters in each paragraph.

(**2**) **store** together all candidate pairs, after having eliminated duplicate paragraphs in *lang1*. Here, we define a hash function as in equation 1 and only insert a new row if there is no previous row with the same hash on *lang1*.

$$hash(P) = \sum_{i=a}^{z} count(i, lowercase(P)) \cdot prime_i \quad (1)$$

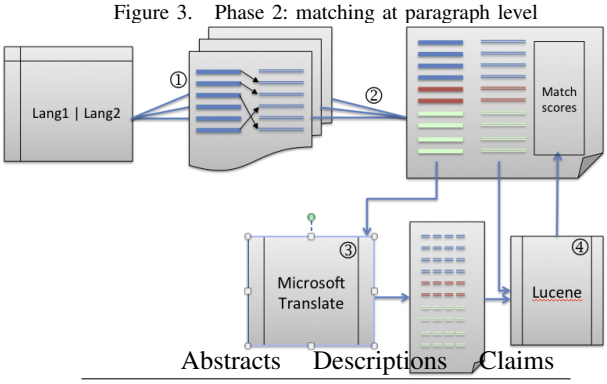
where P is the paragraph at hand, $count(i, lowercase(P))$ is the number of occurrences of character i in the lowercased version of P and $prime_i$ is a prime number associated with character i. The hash simply looks at the alphabet letters, and considers therefore to be a duplicate those paragraphs which have the same amount of each type of letter. This makes sense for this domain, because often paragraphs are extremely similar, but for a misplaced newline or separating line.

(**3**) **translate** the paragraphs from *lang1* to *lang2* using an off-the-shelf translator, in this case, the free version of the Microsoft Translator available as an API from the Windows Azure Marketplace.

(**4**) finally, **index** both the translated results obtained in step 3, and the corresponding pairs from step 2 into one Lucene index and compute a similarity score between the translations and the original pairs. We use the translations as queries, take the first paragraph returned by Lucene and associate the resulting score with the pair (*lang1*, *lang2*) created at step 2.

We need to note that the matches generated at step 2 totally lack semantics and are therefore likely to contain many pairs which are, to a human being, clearly superfluous. However, because the different documents are free (or obliged) to re-order the set of claims or the sections of the descriptions, there is no apriorical filter we can impose, other than assuming that a paragraph of x words cannot be the translation of a paragraph of more than $1.2x$ words in the other language. Even this filter may change as a function of the specific language pair. It is true that an analysis of relative language length is difficult to make, and certainly not the focus here, but the point of the filter is not that, but rather to discard clearly mismatching paragraphs.

Figure 3. Phase 2: matching at paragraph level



	Abstracts	Descriptions	Claims
EN	31788	10107	10417
ES	783	11	12
pairs	2423	10	10

Table II
NUMBER OF SECTIONS IN EN, ES, AND PAIRED (THOUSANDS)

The results obtained at the end of Phase 2 can then be considered as input to an SMT workflow.

III. THE SPANISH USE-CASE

In the remainder of this presentation we instantiate the method described above for extracting English-Spanish comparable paragraphs. Note from Figure 1. and Table I. that Spanish is the most frequent language in the collection, for which no parallel claims, nor descriptions exist.

A. Phase 1

Table II. shows how many abstracts, descriptions and claims were found for each language and the number of resulting pairs. A couple of observations on this table:

a. the numbers of English abstracts, descriptions and claims are smaller than those presented in Figure 1. because we observed that some sections were the result of machine translations. Where indicated as such, we removed them from consideration.

b. the number of pairs of abstracts is larger than the number of Spanish abstracts. This is due to the existence, within the same family, of multiple English abstracts. Each of them will be mapped to the same Spanish abstract. This is desired behaviour, because there is no way to know a priori which is the best match.

At the end of this phase we have a small number of descriptions and claims pairs, each with many paragraphs.

B. Phase 2

In Phase 2 of this test case we show a proof of concept. Here, the main limitation is the translation engine. In our case, we used the free version of the Microsoft Translator, which allows us to translate up to 2 million characters per month. In any other deployment, any other SMT system, trained on a general corpus is usable.

	Descriptions	Claims	Total
EN	656	520	1176
ES	811	325	1136
pairs	8919	4496	13813

Table III
NUMBER OF PARAGRAPHS SELECTED FOR PROCESSING AND ANALYSIS

We tested Phase 2 with 30 claim- and 15 description-pairs, for a total of just under 1 million characters. Abstracts are not the focus of this study because on one hand, there exist already 783'000 pairs in the database and, on the other hand, they have only one paragraph, so there is no real matching to do. For each of the claims and descriptions, we extracted the paragraph tags from each section, and created the table shown in Figure 3. after step 2. The resulting table contains a total of 13'813 pairs, broken down into unique paragraphs and languages as shown in Table III.

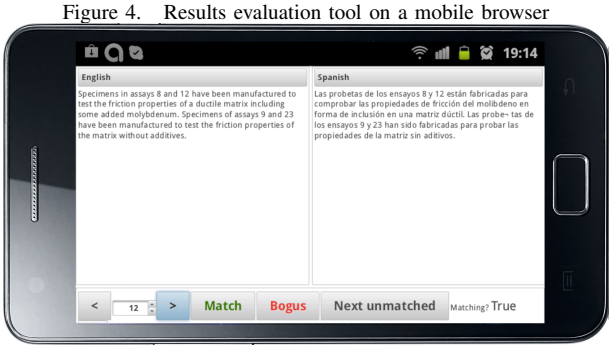
We generated 1176 Spanish translations for the English paragraphs and indexed them together with the rest of the Spanish paragraphs. For each of the translations, we then use the *MoreLikeThis* functionality in Lucene to identify the most similar paragraph from the existing Spanish ones. The matching procedure does not take into account from which document did the matching paragraph come. This adds complexity, but is necessary because it is not always the case that all information will be present in one document. An invention granted one patent in one office may be split into two patents in another office. In such a case, at least the claims will be divided, if not the description as well.

Given that the paragraphs sometime repeat without much change, particularly in the claims sections, it is often the case that the most similar section is found to be one of those that were generated by another translation, rather than one of the existing Spanish paragraphs. We chose as a match to a given English paragraph the Spanish paragraph most similar to its translation, while not itself the result of the translation process. Also, if no such paragraph is found in the top 10 most similar results, then no match is returned. Such situations happened quite frequently and the result was that only a total of 242 English paragraphs (143 from descriptions, 99 from claims) were assigned a match among the Spanish paragraphs.

C. Results analysis

To evaluate the 242 pairs resulting from Phase 2 of our method, a simple web interface was designed and implemented. The interface showed every pair and allowed the user to indicate if the two pairs matched or not. The "matched" judgement did not require the translation to be perfect. A pair was said to be of matching paragraphs if the information expressed therein was the same and the pair could therefore be sent to subsequent alignment tools in preparation for an SMT process. The web interface is easy enough to use on a mobile browser, as shown in Figure 4.

Figure 4. Results evaluation tool on a mobile browser



	Descriptions	Claims	Total
Matching	115	69	184
Not matching	28	30	58
Total	143	99	242

Table IV
RESULTS AFTER THE MANUAL EVALUATION OF THE 242 AUTOMATICALLY IDENTIFIED PAIRS

This significantly facilitates the process. The author of this paper did the evaluation, with basic knowledge of Spanish. The evaluation takes an average of 20 seconds per pair. The results are shown in Table IV.

In what follows, we will go through some of the most frequent examples where we considered that the match was not properly done by the system. As we will see, most of them could actually be corrected with a lower level aligner.

1) Small sizes: The very small paragraphs (10 words or less) in the set of 242 analysed were found to be mismatched. There are two kinds of problems here. First, totally mismatched paragraphs. There are two causes for this: 1. there is too little text to calculate a meaningful similarity value; and 2. the difference in paragraph splitting practice in the different patents makes it such that some small paragraphs are included within other paragraphs in one of the texts, and are therefore not to be found.

Second, there are small claims where just one noun phrase is changed. This happens rather frequently in claims, as it is not uncommon to have lists of claims, with very small differences between each other. For these kinds of mismatches it is difficult to say that they are indeed problematic. However, they are listed here as mismatches because there is the danger that an SMT system will learn the wrong pairs of words, if too many examples like this are found. Ultimately, there is a reason for which the lawyer drafting the claim decided to list the differences explicitly.

2) Similar claims: Most of the mismatched paragraphs are from the claims section. This is not only in absolute terms (30 of 58 mismatched paragraphs were from claims) but even more so in relative terms: 30% of paragraphs from claims were mismatched, compared with only 20% from abstracts. In part, this is explained by the previous observation regarding the sizes, but even for longer claims, it is often the case that they are phrased very similarly, with only minor modifications. In many cases however, it is

not clear where the exact matched/not-matched decision line should stand. Sometimes an element, or a filter is changed or added, other times the claims are very similar but one describes a method while its identified pair the device. This is a significant difference and it would be wrong for the SMT system to learn that the two are the same thing.

3) Faulty paragraph breaks: Both in the claims and descriptions, we encounter among the mismatched pairs, paragraphs that, while sharing a substantial amount of text, are not the same. This is due to the differences in the location of the paragraph tags in the XML document. Sometimes, the paragraph tags seem to be simply misplaced, most likely due to an automated process at some point in the life time of the document.

4) Other observations: In doing the manual evaluation of the matched paragraphs it has not always been easy understanding even the English version. To some extent, one wonders to what extent this is explained by the *"patentese"* or whether it is possible that somewhere in the life-cycle of the document, machine translation was applied and not identified as such. A full study, by trained experts is unfortunately outside the possibilities of this paper.

IV. Related Work

There is relatively little work on extracting comparable corpora from patent documents. Many components of this process are, on the other hand, part of the general scientific knowledge. The use of the number of words as a feature in sentence alignment was introduced by [6], while Gale and Church preferred the use of characters [7]. To what extent these observations apply in this particular context of patent data is still to be investigated. [8] found that the character length ratios is similar in different kinds of texts, but patents were not part of the study. More recently, this method has also been used on Asian languages [9].

The use of intermediate translations, or dictionaries, is also in practice since the early 90s. [10], using a dictionary to guide the matching process, extended a previous method introduced by [11]. [12] uses a simple translation model, which does not take into account word order, to improve upon the results of Gale and Church mentioned above.

The most similar work is that of [13], focusing on Asian languages. The difference lies mainly in the focus. While the method presented here is aimed at preparing a corpus of comparable data for further processing, their work takes a more restricted set of patents as the input (only patents filed via WIPO, the so-called "PCT route"), but looks deeper, at sentence level.

V. Conclusions

This work takes existing components and data, and combines them together to obtain a new multilingual corpus of comparable data. It starts from the observation that, while the patent system is fundamentally a multilingual one, the set of multilingual patents is actually very small. Instead, the same inventions are published by different offices separately, in their respective languages, and the only way to link them with any measure of reliability is via their priority numbers, by grouping them into so-called patent families.

The use of thus linked patents in an SMT system is difficult because the descriptions and claims may be too different between the different instances of the invention disclosure. To make this manageable, we split them into paragraphs and use a public translation system to pair them.

We present a case study for English-Spanish patents, provide a manual evaluation system and results, and observe the most frequent causes of mistaken match between paragraphs. Most of them are likely to be filtered out in subsequent stages of sentence alignment, as they are the result of either broken paragraph tags, or the similar style of writing claims, where most of the terms repeat and only a small part changes at every claim.

Acknowledgements

Mihai Lupu is partially supported by the PLuTO (ICT-PSP-250416), PROMISE (NoE-258191) and IMPEx (FFG-825846) projects. Acknowledgments go to Fairview Research, and to patent experts such as H. Thomas, T. Loughbrough and S. Adams, for insights into the patent system.

References

[1] K. H. Atkinson, "Towards a more rational patent search paradigm," in *Proc. of PaIR*, 2008.

[2] J. Tinsley and P. Sheridan, "Pluto annual public report," http://bit.ly/JPdXKd, 2011.

[3] Fairview Research, "Alexandria patent data warehouse," http://www.intellogist.com/wiki/Alexandria, 2011.

[4] S. Adams, "The text, the full text and nothing but the text: Part 1 - standards for creating textual information in patent documents and general search implications," *WPI Journal*, vol. 32, no. 1, 2010.

[5] N. Oostdijk, E. D'hondt, H. van Halteren, and S. Verberne, "Genre and domain in patent texts," in *Proc. of PaIR*, 2010.

[6] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. Mercer, "Word -sense disambiguation using statistical methods." *ACL*, vol. 29, 1991.

[7] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," *Computational Linguistics*, vol. 19, no. 3, 1993.

[8] J. Véronis, *Ingénierie des langues, Traité IC2-Série Informatique et SI*. Èditions Hermes Science, Paris, 2000, ch. 6 : Alignement de corpus multilingue.

[9] H. Ding, L. Quan, and H. Qi, "The Chinese-English bilingual sentence alignment based on length," in *Proc. of IALP*, 2011.

[10] F. Debili and E. Sammouda, "Appariement des phrases de textes bilingues français-anglais et français-arabe." in *Proc. of COLING*, 1992.

[11] M. Kay and M. Röscheisen, "Text-translation alignment," Xerox Palo Alto Research Center, Tech. Rep., 1988.

[12] S. F. Chen, "Aligning sentences in bilingual corpora using lexical information." *ACL*, vol. 31, no. 9-16, 1993.

[13] B. Lu, B. K. Tsou, T. Jiang, O. Y. Kwong, and J. Zhu, "Mining large-scale parallel corpora from multilingual patents: An english-chinese example and its application to smt," in *Proc. of CIPS-SIGHAN*, 2010.