Using partly multilingual patents to support research on multilingual IR by building translation memories and MT systems

Lingxiao WANG[1,2], Christian Boitet[2], Valérie Bellynck[2], Mathieu Mangeot[2]

[1] SAS Lingua et Machina, c/o Inria, Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay Cedex, France
[2]LIG-GETALP, Bâtiment IM^{2}AG B, Laboratoire LIG, 41 rue des mathématiques
38400 ST Martin d'Hères, France
E-mail: {*Lingxiao.Wang, Christian.Boitet, Valerie.Bellynck, Mathieu.Mangeot*}@imag.fr

Abstract

In this paper, we describe the extraction of directional translation memories (TMs) from a partly multilingual corpus of patent documents, namely the CLEF-IP collection, and the subsequent production and gradual improvement of MT systems for the associated sublanguages (one for each language), the motivation being to support the work of researchers of the MUMIA community. First, we analysed the structure of patent documents in this collection, and extracted multilingual parallel segments (English-German, English-French, and French-German) from it, taking care to identify the source language, as well as monolingual segments. Then we used the extracted TMs to construct statistical machine translation systems (SMT). In order to get more parallel segments, we also imported monolingual segments into our post-editing system, and post-edited them with the help of SMT.

Keywords: Extraction of parallel segments, SMT, CLEF patent collection, translation memories, source langue identification, support for CLIR

1. Introduction

Parallel corpora have an important role in the natural language processing (NLP), and are a valuable resource for many NLP applications, such as statistical machine translation (SMT), cross-lingual information retrieval and multilingual lexicography. Patent description documents, because they often contain multilingual translations of some segments, are also seen as an important source of parallel corpora. Much work has been done on this topic, such as (Utiyama and Isahara. 2007), (Lu et al., 2009), and (Wäschle and Riezler, 2012).

In this paper, we describe our method for extracting a multilingual parallel corpus from a patent corpus, namely the CLEF-IP collection[1], and present how to use these data. From the extracted multilingual parallel segments (English-German, English-French, and French-German), we built a translation memory (TM) and added it into our iMAG/SECTra system (Wang and Boitet, 2013). We then produced several SMT systems from this MT. In order to contribute to WG2 of the MUMIA[2] community on infrastructure, we transformed the collection of patents in a website where each patent is monolingual, and can be accessed (and collaboratively) post-edited into any language, using the above desired MT system when applicable, and free MT Web servers otherwise (e.g., for access in Chinese).

2. The CLEF-IP Collection

The latest version of collection corpus is the same as the one used in the CLEF-IP 2011 lab (the data corpus used in 2012 and 2013 is the same as the one used in 2011), so our work is based on the CLEF-IP 2011 collection. This collection comprises more than 117 GB of multilingual patent documents derived from European Patent Office (EPO) and World Intellectual Property Organization (WIPO) sources. The CLEF-IP 2011 collection is composed of about 3.5 M XML files containing the textual part (no images) of about 1.5 M partially multilingual patent documents, corresponding to over 1.5 million patents published until 2002.

A patent document of the CLEF-IP 2011 collection is an *application document*, a *search report*, or a *granted patent document*, which is stored as a XML file. Each patent document has a unique patent name (EP for the EPO, or WO for the WIPO, followed by a series of digits and a code A or B[3], like EP-0071719-B1.xml). Different information and different content of the patent document are stored in various XML fields, such as <*bibliographic-data*>, <*invention-title*>, <*abstract*>, <*description*>, <*claims*>, <*copyright*>, etc., and the fields of some patent documents also have subfields. The content of the various XML fields can be in English, French, or German (official languages of the EPO). However, not all segments of patent documents have content in these fields.

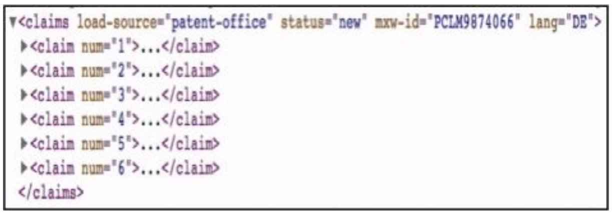
Each XML patent document of CLEF-IP 2011 has an associated document language, which we can find it in the <*patent-document*> field. During our extraction process, we consider the document language as the source language. We analyzed patents with respect to the structure of their XML fields, and found that four main fields may have parallel segments: <*invention-title*>, <*abstract*>, <*description*>, and <*claims*>. Each field may have some subfields, for example, a field <*claims*> may contain 6 <*claim*> subfields in EP-0260000-B1.xml

[1] Cross-Language Experiment Forum (CLEF), http://www.clef-campaign.org,
and http://www.ifs.tuwien.ac.at/~clef-ip/index.html

[2] MUMIA (MUltilingual, multimodal, Multifaceted Information access is a COST action (CE1002) of the UE. Many members of its network do research on CLIR in patents.

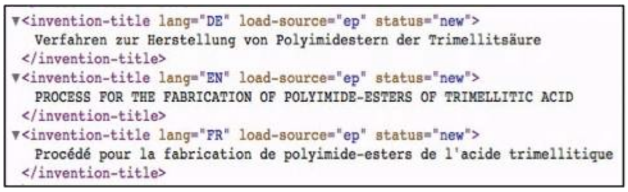
[3] List of patent document kind codes: https://register.epo.org/help?topic=kindcodes and http://www.wipo.int/patentscope/en/wo_publication_information/kind_codes.html

(Figure 1). We begin with these fields, looking for fields that appear more than once in the patent document and each field with a different language attribute. For example, Figure 2 shows an <*invention-title*> field with 3 different language attributes (lang="DE", lang="EN", and lang="FR"). Each field also contains some content, in the language that corresponds to its attribute.



```
▼<claims load-source="patent-office" status="new" mxw-id="PCLM9874066" lang="DE">
  ▶<claim num="1">...</claim>
  ▶<claim num="2">...</claim>
  ▶<claim num="3">...</claim>
  ▶<claim num="4">...</claim>
  ▶<claim num="5">...</claim>
  ▶<claim num="6">...</claim>
 </claims>
```

Figure 1: <*claims*> has 6 <*claim*> subfields in EP-0260000-B1.xml



```
▼<invention-title lang="DE" load-source="ep" status="new">
   Verfahren zur Herstellung von Polyimidestern der Trimellitsäure
 </invention-title>
▼<invention-title lang="EN" load-source="ep" status="new">
   PROCESS FOR THE FABRICATION OF POLYIMIDE-ESTERS OF TRIMELLITIC ACID
 </invention-title>
▼<invention-title lang="FR" load-source="ep" status="new">
   Procédé pour la fabrication de polyimide-esters de l'acide trimellitique
 </invention-title>
```

Figure 2: Example of an <*invention-title*> field with 3 different language attributes and the corresponding contents in 3 different languages

3. Extraction of Parallel Data

We started from the 3.5 million XML files corresponding to 1.5 million patents. The first goal was to extract from them as many useful parallel segments as possible. First, we traverse every patent document. For each patent document, we select the source language from the <*patent-document*> field, according to the language attribute of this field. Second, we search the parallel segments contained in the four main fields (<*invention-title*>, <*abstract*>, <*description*>, and <*claims*>). Sometimes, some fields occur with different language attribute than the document language. For example, in *EP-0260700-B1.xml*, English is the document language, but <*claims*> segments do not exist in English, only German and French versions are available. Even though it is always desirable to collect as much text as possible, it is even more important to ensure the quality of the texts, so in this case, we do not store the German and French parts as a parallel segment.

All fields, which appear more than once in a patent document and have different language attributes, are treated as a collection. In general, an EPO patent document has a maximum of 3 languages (English, French, and German). We chose as source segment the segments whose language attribute is consistent with the source language, and then extract the target parallel segments from the other fields. For example, in EP-0301015-B1.xml, the source language is English, and the <*claims*> field appears 3 times. Hence, we use the English part of the claims fields as the source segments, and consider the French and German parts as the target segments. The source segment and the target segments are then stored separately into different files. In the above example, the source segment has been stored into *CLEF_claims_en-fr.en* and *CLEF_claims_en-de.en*, and the target segments in *CLEF_claims_en-fr.fr* and *CLEF_claims_en-de.de,* respectively. In order to reduce the noise in the data, we keep only the extracted text, and remove all tags.

Not all the extracted data is fully suitable for direct use for NLP applications. We have to clean the extracted data and eliminate some noise. First, we split the text into sentences, and then remove useless whitespace, and duplicate sentences. For alignment, we use the LF Aligner[4], an open-source tool based on Hunaligne (Varga et al., 2005), which has the widest linguistic backbone (a total of 32 languages), and permits the automatic generation of dictionaries in any combination of these languages. Aligned segments are prepared bilingually for 4 types (title, abstract, description, and claims), and all 6-language pairs (de_en, de_fr, en_de, en_fr, fr_de, fr_en).

4. Some Statistics About the Corpus

Table 1 shows the number of segments and words that are extracted from the title and claims fields on the source and the target after segment aligning. All extracted parallel sentences are saved in TMX and TXT formats, and can be found at http://membres-liglab.imag.fr/wang/downloads

5. Application in SMT

We used our extracted parallel corpus (the title and claims fields) to construct SMT systems with the Moses Toolkit (Koehn et al., 2007). First, preparing the development set and the test set, we extracted 2,000 sentences for training the feature weights of Moses, and extracted 1,000 sentences for testing. Then we use the rest to train translation models of Moses. We actually built SMT system for only 3 directions: de-en, de-fr, and en-fr.

The systems also include 5-gram language models trained on the target side of corresponding parallel texts using IRSTLM (Federico et al., 2008). The feature weights required by the Moses decoder were further determined with MERT (Och, 2003) by optimizing BLEU scores on the development set (1,000 sentences). The test sets were translated by the resulting systems and then used to evaluate the systems in terms of BLEU scores (Papineni et al., 2001), as shown in Table 2.

6. Post-editing Monolingual Sentences Pre-translated by SMT

When we extracted parallel sentences from the CLEF-IP collection, we also derived large amount of monolingual sentences, which are not translated in the patent documents. The language of patents, although having a large amount of vocabulary and richness of grammatical structure, can be considered as a specialized sub language, because its grammar is quite restricted compared to that of the whole language. Second, patents have attributes of domain, this has been proven in some works, for example, (Wäschle and Riezler, 2012). Third, recent experiments in specializing empirical MT systems have shown that remarkably good MT results can be obtained (Rubino et al., 2012). So we combine these features with framework iMAG/SECTra (Wang and Boitet, 2013).

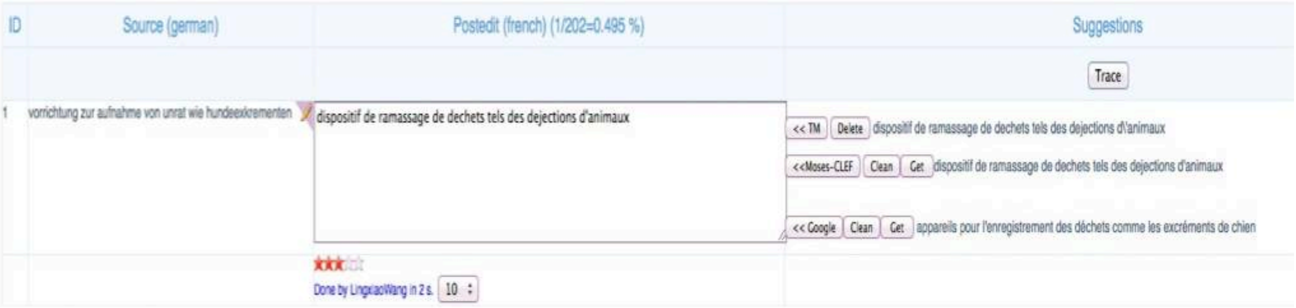
[4] http://sourceforge.net/projects/aligner/

Language pairs		Title		Claims	
		Segments	Words	Segments	Words
de-en	de	311,298	2,038,785	1,696,498	62 M
	en		2,582,703		71 M
de-fr	de	311,184	2,036,112	1,661,419	79 M
	fr		2,482,257		86 M
en-de	en	884,759	6,661,481	5,218,024	332 M
	de		5,508,289		296 M
en-fr	en	884,727	6,661,322	5,373,452	330 M
	fr		8,538,012		380 M
fr-de	fr	106,211	963,508	572,356	36 M
	de		1,204,439		37 M
fr-en	fr	106,246	1,285,467	586,498	38 M
	en		1,048,374		37 M

Table 1: Number of extracted segments as source and target after segment aligning in the <*title*> and <*claims*> fields

Language pairs	Development set	Test Set
de-en	37.46	31.67
de-fr	35.41	28.72
en-de	43.16	36.01
en-fr	42.59	38.82
fr-en	44.12	42.61
fr-de	34.85	30.14

Table 2: BLEU scores of SMT systems



ID	Source (german)	Postedit (french) (1/202=0.495 %)	Suggestions
			Trace
1	vorrichtung zur aufnahme von unrat wie hundeexkrementen	dispositif de ramassage de dechets tels des dejections d'animaux	<< TM Delete dispositif de ramassage de dechets tels des dejections d\'animaux <<Moses-CLEF Clean Get dispositif de ramassage de dechets tels des dejections d'animaux << Google Clean Get appareils pour l'enregistrement des déchets comme les excréments de chien
		★★★☆☆ Done by LingxiaoWang in 2 s. 10	

Figure 3: Interfaces of post-edting on SECTra

We store all monolingual sentences into html files, and add them into iMAG/SECTra. Pre-translation is provided by SMT systems, which are built with data extracted from the CLEF-IP 2011 collection. Figure 3 presents an example, where source sentences (de) are pre-translated (fr) by Moses and Google.

Figure 3 shows SECTra translation editor interface, similar to those of translation aids and commercial MT systems. It makes post-editing much faster than in the presentation context. Not yet post-edited segments can be selected, and global search-and-replace is available. All post-edited sentences are saved in a translation memory called CLEF-IP. When it becomes large enough after some period of using SECTra (about 10-15000 'good' bi-segments for the sublanguages of classical web sites), it can be used to build an empirical MT system for that sublanguage, and then to improve it incrementally as time goes and new segments are post-edited.

iMAG/SECTra also provides more languages options for patent translation, such as Chinese, Hindi, or Arabic, using SMT or some online free web servers such as Google Translator, Systran, or Bing.

7. Support research on multilingual IR

Multilingual information search becomes important due to the growing amount of online information available in non-English languages and the rise of multilingual document collections. Query translation for CLIR became the most widely used technique to access documents in a different language from the query. For query translation, SMT is one way in which those powerful capabilities can be used (Oard, 1998).
Our 3 SMT systems offer translation service by API. IR systems can use them directly. Due to robustness across domains and strong performance in translating named entities (like titles or short names), using SMT for CLIR can produce good results (Kürsten et al., 2009).

8. Conclusion and future work

In this paper, we gave an account of the extraction of a multilingual parallel corpus from the CLEF-IP 2011 collection. We first analyzed the structure of the patent documents of this collection and chose the fields to be extracted. To ensure the quality of parallel data, we cleaned them and aligned them with LF Aligner. The first version of the extracted patent parallel corpus consists of 3 languages, 6-language pairs, and is available in different formats (plain text files for Moses and TMX). This corpus is available to the research community. We also developed 3 specialized Moses-based SMT systems, from the TM resulting from the extraction process, and evaluated them, setting good BLEU scores on segments for which no translation was presented in the CLEF-IP 2011files. We also transformed the initial collection of multilingual files into 3 collections of monolingual files, keeping only the source language text in each segment, and accessible in many languages using 3 dedicated iMAGs, and using the TM extracted from the original multilingual files. Multilingual access is provided by using our 3 Moses systems for the 3 corresponding language pairs, and other online free MT systems for the other language pairs.

One interesting perspective is the development of an infrastructure for the multilingual aspect of MUMIA-related research on patents. In the near future, we will setup a web service to support evaluation of the translation quality, both subjective (based on human judgments) and objective (task-related, such as post-editing time, or understanding time).

What has been done so far should enable researchers on CLIR applied to patents to include the multilingual aspect in their experiments. In future experiment, we plan to ask visitors of 3 websites to post-edit the MT "pre-translations". Interactive post-editing will transform the MT pre-translations of segments having no translation in the original CLEF-IP 2011 corpus into good translations, and the SMT systems will thus be incrementally improvable.

9. References

Oard, Douglas W. (1998). A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval, *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup,* p.472-483, October 28-31, 1998.

Eisele, A., and Yu C. (2010). "MultiUN: A Multilingual Corpus from United Nation Documents". *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Federico, M., Bertoldi, N., and Cettolo, M. (2008), "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models", *Proceedings of Interspeech*, Brisbane, Australia, 2008.

Kürsten, J., Wilhelm, T., and Eibl, M. (2009). The Xtrieval framework at CLEF 2008: domain-specific track. *Proceedings of CLEF,* pages 215–218, 2009.

Koehn, P. (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". *Proceedings of Machine Translation Summit X*, Phuket, Thailand

Koehn, P., Hoang, H., Birch, A., Callison-Birch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "Moses: Open source toolkit for statistical machine translation". *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic.

Lu, B., Tsou, B.K., Zhu, J., Jiang, T. and Kwong, O.Y. (2009). "The construction of a Chinese-English patent parallel corpus". *Proceedings of the MT Summit XII*, Ottawa, Canada 2009.

Och, F.J. (2003). Minimum error rate training in statistical machine translation, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Sapporo, Japan, Volume 1, 160-167.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *In Proceedings of the Association of Computational Linguistics*, pp. 311–318.

Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC'2006). Genoa, Italy, 24-26 May 2006.

Rubino, R., Huet, S., Lefèvre, F., and Linarès., G. (2012). Post-édition statistique pour l'adaptation aux domaines de spécialité en traduction automatique, *In Conférence en Traitement Automatique des Langues Naturelles,* pp. 527-534, Grenoble, France.

Utiyama, M., and Isahara, H. (2007). A Japanese-English Patent Parallel Corpus. *Proceedings of MT Summit XI*.

Varga, D., Halacsy, P., and et al. (2005). Parallel Corpora for Medium Density Languages. *RANLP 2005 Conference*.

Wäschle, K., and Stefan, R. (2012). Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Proceedings of the 5th Information Retrieval Facility Conference, IRFC 2012*, Vienna, Austria, July 2-3, 2012 12–27.

Wang, L., and Boitet, C. (2013), Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost. *Proceedings of MT Summit XIV, The 2nd Workshop on Post-Editing Technologies and Practice.* Nice, France 2 - 6 September 2013.