COPPA V2.0: Corpus Of Parallel Patent Applications
Building Large Parallel Corpora with GNU Make

Marcin Junczys-Dowmunt, Bruno Pouliquen, Christophe Mazenc
World Intellectual Property Organization
34, chemin des Colombettes
CH-1211 Geneva 20

{Marcin.Junczys-Dowmunt, Bruno.Pouliquen, Christophe.Mazenc}@wipo.int

Abstract

WIPO seeks to help users and researchers to overcome the language barrier when searching patents published in different languages. Having collected a big multilingual corpus of translated patent applications, WIPO decided to share this corpus in a product called COPPA (Corpus Of Parallel Patent Applications) to stimulate research in Machine Translation and in language tools for patent texts. A first version was released in 2011 but contained only French and English languages. It has been decided to release a major update of this product containing newer data (from 2011 up to 2014) but also other languages (German, English, French, Japanese, Korean, Portuguese, Spanish, Russian and Chinese). This corpus can be used for terminology extraction, cross-language information retrieval or statistical machine translation. With the new version a huge number of files (more than 26 million) has to be processed. We describe the technical process in details.

Keywords: Parallel Corpus of Patents, Build System, GNU Make

1. Introduction

WIPO is a specialized agency of the United Nations dealing with Intellectual Property. WIPO notably administers the Patent Cooperation Treaty (PCT[1]) and while publishing international patent applications, translates the associated titles and abstracts into both English and French. These applications are submitted in one of the PCT publication language (Arabic, German, English, French, Japanese, Korean, Portuguese, Spanish, Russian, or Chinese). Therefore WIPO has an extensive parallel corpus of manually translated patent documents collected over time, especially for the language pair English-French (more than 1.7 million documents), but also from/to other languages (German, Japanese, Korean, Portuguese, Spanish, Russian, or Chinese[2]).

PCT Patent applications are published on the PATENTSCOPE search engine[3], together with other national and international collections. WIPO has investigated techniques for overcoming the language barrier such as cross-language retrieval and machine translation, and developed its own tools based on the open-source toolkit Moses (Koehn et al., 2007), benefiting from academic research results in machine translation.

Cross language Information Retrieval: The fact that WIPO has searchable patent documents in various languages has led to building a tool (called CLIR[4]) to allow users to easily search simultaneously in those various languages.

Statistical Machine Translation: The COPPA corpus has first been fed into an open-source-based statistical machine translation tool (called TAPTA: Translation Assistant for Patent Titles and Abstracts[5]). It can translate texts from English into German, French, Japanese, Korean, Spanish, Russian or Chinese, and vice-versa, (Pouliquen et al., 2011).

In order to further promote research in this field, WIPO decided in 2011 to release the PCT parallel English-French corpus in an easy-to-use TMX format in a product called COPPA (Pouliquen and Mazenc, 2011). However this corpus contained only English and French texts, and it has been decided to extend the corpus with more languages and more recent applications.

2. COPPA: Corpus Of Parallel Patent Applications

The segments included in the corpus are obtained by aligning the sentences of the abstracts and titles of published PCT applications with their translations, the translations having been produced by professional patent translators (More than 200,000 new PCT applications are published every year). It is therefore a gold mine for linguistic research such as terminology extraction, translation memory building and research on Machine Translation.

With the goal of supporting innovation in the Machine Translation field, WIPO offers the updated corpus under the same conditions as before, the product being notably free of charge for academic and private research institutions for research purposes only; in return those institutions commit to share their published results with WIPO.

WIPO hopes that the wide availability of this improved corpus will actively contribute to progress in building more accurate machine translation systems for patent texts with the

[1] Also called PCT application, see WIPO (2010).

[2] Only 25 PCT applications were published in the Arabic language (9/10/2015). We decided for this version not to include them.

[3] http://www.wipo.int/patentscope/search

[4] Publicly available at: https://patentscope.wipo.int/search/en/clir/clir.jsf

[5] Publicly available at https://www3.wipo.int/patentscope/translate

Language pair	Documents	Sentences	Tokens	Characters
en-de	289’287	982’510	36’814’520	225’972’826
en-es	18’303	62’057	2’328’713	14’624’745
en-fr	2’570’292	10’557’032	316’271’950	2’006’750’520
en-ja	312’664	1’036’614	42’127’479	264’578’974
en-ko	41’093	120’534	5’813’474	37’047’347
en-pt	2’001	7’000	261’843	1’696’039
en-ru	6’972	37’261	1’241’791	7’841’040
en-zh	289’287	982’510	36’814’520	225’972’826
Total	3’240’612	12’803’008	404’859’770	2’558’511’491

Table 1: Statistics for the complete corpus. The total does not reflect unique documents as all the documents are available in English and French (a Japanese document - in the en-ja corpus - will also be part of the en-fr subcorpus)

Language		Into English	From English
German	(de)	44.68	30.85
Spanish	(es)	32.97	34.27
French	(fr)	51.06	51.74
Japanese	(jp)	30.54	25.84
Korean	(ko)	25.99	27.95
Russian	(ru)	24.48	32.37
Chinese	(zh)	35.77	32.68

Table 2: BLEU scores for SMT output with the provided test set

ultimate goal of lowering the linguistic barrier for inventors and the general public and of improving the efficiency and the accessibility of the international patent system.

2.1. Statistics

The corpus now contains more than 300 Million words (English-French), for comparison (only for English-French), the previous COPPA version contained 180 Million words, the European corpora (DGT-Acquis/DCEP, (Steinberger et al., 2006)) are about 100 Million words each. See Table 1 for full details.

2.2. Usage in statistical machine translation

We trained our “TAPTA” software on the data provided. The evaluation results are summarized in table 2 (note that the Portuguese COPPA data is too small and has been ignored).

For each language, the new corpus is divided into three distinct sets: a training set (all data until 2014), a development set, and a test set (data taken from early 2015 applications). The training of any statistical models should be done exclusively on the given training set.

Sentences longer than 80 words were discarded. To speed up the word alignment procedure, we split the training corpora into four equally sized parts that are aligned with MGIZA++ (Gao and Vogel, 2008), running 5 iterations of Model 1 and the HMM model on each part.[6] We use a 5-gram language model trained from the target parallel data, with 3-grams or higher order being pruned if they occur only once. Apart from the default configuration with a lexical reordering model, we add a 5-gram operation sequence model (Durrani et al., 2013) (all n-grams pruned if they occur only once) and a 9-gram word-class language model with word-classes produced by word2vec (Mikolov et al., 2013) (3-grams and 4-grams are pruned if they occur only once, 5-grams and 6-grams if they occur only twice, etc.), both trained using KenLM (Heafield et al., 2013). To reduce the phrase-table size, we apply significance pruning (Johnson et al., 2007) and use the compact phrase-table and reordering data structures (Junczys-Dowmunt, 2012). During decoding, we use the cube-pruning algorithm with stack size and cube-pruning pop limits of 1,000.

The development set has been used to tune Moses parameters (using MERT) for the obtained model, while the test set has been used to measure the BLEU scores of the final model. As a result, research teams can use the COPPA corpus in the same conditions, and have a first baseline to benchmark their solution against the BLEU scores obtained by WIPO.

2.3. Technical details

The previous version of COPPA was using the widely used TMX format[7], however we found it more convenient to use TEI[8] for this version and use scripts to export from this format to others. Each document contains, in addition, some meta data that can be extremely useful to use for machine learning: the associated International Patent Classification codes (IPC codes) (which can be used to train “domain-aware” tools as with CLIR and TAPTA),the main applicant’s name, the language of filing (which is a good indication on the direction the translation was done), the application identifier (which also contains the patent office identification) and two dates (application date and publication date).

2.4. Availability

The corpus is available for free for research purposes and for a nominal fee for other purposes, order form and details are available at: `http://www.wipo.int/patentscope/en/data/products.html#coppa`

[6] We confirmed that there seemed to be no quality loss due to splitting and limiting the iterations to simpler alignment models.

[7] `http://www.lisa.org/tmx`

[8] `http://www.tei-c.org`

```
<?xml version="1.0" encoding="utf-8"?>
<TEI.2 id="WO2014071330-fr" lang="fr">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>PROCÉDÉ ET SYSTÈME DE TRAITEMENT DE LANGA
      </titleStmt>
    </fileDesc>
    <notesStmt>
      <note type="ID">WO2014071330</note>
      <note type="AD">20131105</note>
      <note type="ANID">US2013068360</note>
      <note type="DP">20140509</note>
      <note type="IC">G06F 17/28</note>
      <note type="LGF">EN</note>
      <note type="OF">WO</note>
      <note type="PA">FIDO LABS INC.</note>
    </notesStmt>
  </teiHeader>
  <text>
    <body>
      <head id="1" lang="fr">PROCÉDÉ ET SYSTÈME DE TRAIT
      <div type="abstract">
        <p id="2">
          <s id="2:1" lang="fr">La présente invention co
          <s id="2:2" lang="fr">Des modes de réalisation
          <s id="2:3" lang="fr">Pour accroître la précis
          <s id="2:4" lang="fr">Des modes de réalisation
          <s id="2:5" lang="fr">La sortie du LD est faci
          <s id="2:6" lang="fr">La présente invention co
          <s id="2:7" lang="fr">La présente invention co
        </p>
      </div>
    </body>
  </text>
</TEI.2>
```

(a) French example document

```
<?xml version="1.0" encoding="utf-8"?>
<TEI.2 id="WO2014071330-en" lang="en">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>NATURAL LANGUAGE PROCESSING SYSTEM AND ME
      </titleStmt>
    </fileDesc>
    <notesStmt>
      <note type="ID">WO2014071330</note>
      <note type="AD">20131105</note>
      <note type="ANID">US2013068360</note>
      <note type="DP">20140509</note>
      <note type="IC">G06F 17/28</note>
      <note type="LGF">EN</note>
      <note type="OF">WO</note>
      <note type="PA">FIDO LABS INC.</note>
    </notesStmt>
  </teiHeader>
  <text>
    <body>
      <head id="1" lang="en">NATURAL LANGUAGE PROCESSING
      <div type="abstract">
        <p id="2">
          <s id="2:1" lang="en">A natural language proce
          <s id="2:2" lang="en">Embodiments of the NLP s
          <s id="2:3" lang="en">Rules can be added or mo
        </p>
      </div>
    </body>
  </text>
</TEI.2>
```

(b) English example document

```
<linkGrp fromDoc="Xml/fr/WO2014/07/13/WO2014071330.xml"
  toDoc="Xml/en/WO2014/07/13/WO2014071330.xml"
  score="0.158818">
  <link type="1-1" xtargets="1;1" score="1" />
  <link type="1-1" xtargets="2:1;2:1" score="0.239642"/>
  <link type="1-1" xtargets="2:2;2:2" score="0.345575"/>
  <link type="1-1" xtargets="2:3;2:3" score="0.526508"/>
  <link type="0-1" xtargets=";2:4" score="0" />
  <link type="0-1" xtargets=";2:5" score="0" />
  <link type="0-1" xtargets=";2:6" score="0" />
  <link type="0-1" xtargets=";2:7" score="0" />
</linkGrp>
```

(c) Sentence alignment information between two documents

Figure 1: TEI-based XML format of corpus files

3. Creating the Parallel Corpus

During processing, we differentiate between primary and secondary language pairs. Primary language pairs consist of one Non-English language and English. Secondary language pairs are formed from all Non-English languages.

Figure 2 illustrates all processing steps for creating the sentence alignment link file from two parallel documents for a primary language pair, here English-French. The shown dependency graph is modeled very closely after our pipeline based on GNU Make.

After converting binary formats (MS Word, WordPerfect) to the presented TEI-XML format, sentence splitting[9] is applied to the XML-file, retaining the original paragraph structure as shown in Figure 1a.

To ensure a high sentence alignment quality, we rely on a two-step approach similar to (Sennrich and Volk, 2011). French documents are translated into English first. We randomly select a subset of 10,000 document pairs and align them using Hunalign (Varga et al., 2005), selecting only 1-1 alignments that are themselves surrounded by 1-1 alignments. This small lower-quality parallel corpus is used to train an SMT system with Moses (Koehn et al., 2007). Following (Sennrich and Volk, 2011) we use significance pruning (Johnson et al., 2007) to filter out noise resulting from alignment errors.

Next, our monolingual sentence aligner BLEU-Champ[10] is applied. BLEU-Champ relies on smoothed sentence level BLEU-2 as a similarity metric between sentences and uses the Champollion algorithm (Ma, 2006) with that metric. To avoid computational bottlenecks for long documents, first

[9]Using Eserix, an SRX-based sentence splitter `https://github/emjotde/eserix`. The algorithm and rules have been extracted from Psi-Toolkit (Graliński et al., 2012).

[10]`https://github/emjotde/bleu-champ`

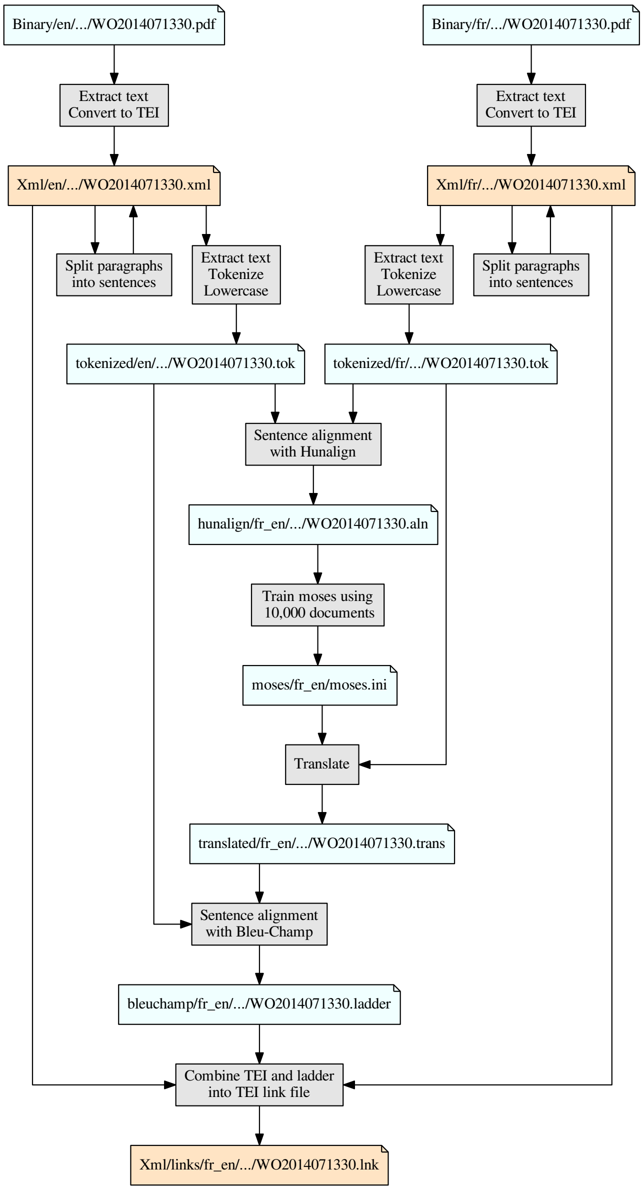

Figure 2: GNU Make dependencies for sentence alignment procedure

a path consisting only of 0-1, 1-0, 1-1 alignments is calculated. In a second step, the search is restricted to a 10-sentence-wide corridor around the best path allowing for all alignment combinations up to 4-4 alignments. This procedure avoids search errors and is fast enough to use the Champollion algorithm with documents consisting of thousands of sentences. Given the English tokenized text and the translated French text, BLEU-Champ produces a ladder file (Hunalign's numeric sentence alignment format) which in the end is combined with the two TEI documents to form the final TEI sentence alignment file (see 1c).

The beige-colored TEI files in Figure 2 are distributed as part of the corpus. Since the link files contain pointers to the original XML documents any set of link files can be used to produce plain-text parallel corpora.

In case of secondary language pairs, the steps are the same with the exception that both documents are translated into English and sentence alignment is performed on the English translation results of both files.

The entire process creates 9,373,728 XML files (document files and link files) meant for distribution and 17,065,732 temporary intermediate targets (plain text tokenized, translated files). Thanks to the use of GNU Make, we can parallelize the processing across 64 physical cores taking advantage of the full available computational power of the used machine. Occasional crashes or interruptions are no problem as the system can easily resume work with minimal overhead.

4. Conclusions

One of the mandates of WIPO is to facilitate access to technical knowledge and information. To achieve this goal, WIPO encourages innovation by providing its corpus of translated patent application (COPPA) free of charge for research purposes.

Our baselines and test sets can serve as reference data for future publications and we would like researchers to explore machine translation techniques beyond the phrase-based approach that was used to produce them. The meta-information and preserved document structure provided can help to advance recent work in document-level translation.

By choosing GNU Make as a build system for our corpus, we created a self-updating processing chain that allows us to easily add new documents with optimal processing steps. By this we can maintain current versions of the corpus and prepare them with minimal effort for possible future updates. The automatic parallelization of GNU Make made it possible to process millions of files in a relatively short time.

5. References

Durrani, N., Fraser, A., Schmid, H., Hoang, H., and Koehn, P. (2013). Can Markov models over minimal translation units help phrase-based SMT? In *ACL*, pages 399–405. The Association for Computer Linguistics.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. ACL.

Graliński, F., Jassem, K., and Junczys-Dowmunt, M. (2012). PSI-Toolkit: Natural Language Processing Pipeline. *Computational Linguistics - Applications*, pages 27–39.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 690–696.

Johnson, J. H., Martin, J., Forst, G., and Kuhn, R. (2007). Improving translation quality by discarding most of the phrasetable. In *In Proceedings of EMNLP-CoNLL?07*, pages 967–975.

Junczys-Dowmunt, M. (2012). Phrasal Rank-Encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *Prague Bull. Math. Linguistics*, 98:63–74.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, 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. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Stroudsburg, USA. Association for Computational Linguistics.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *In Proceedings of LREC-2006*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Pouliquen, B. and Mazenc, C. (2011). COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent barrier at WIPO. In *Proc of MT-Summit XIII*, pages 24–30, Xiamen, China.

Pouliquen, B., Mazenc, C., and Iorio, A. (2011). Tapta: A user-driven translation system for patent documents based on domain-aware statistical machine translation. In Mikel L. Forcada, et al., editors, *Proceedings of th 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 5–12.

Sennrich, R. and Volk, M. (2011). Iterative, mt-based sentence alignment of parallel texts. *18th Nordic Conference of Computational Linguistics, NODALIDA*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., and Tufiş, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.