JaParaPat: A Large-Scale Japanese-English Parallel Patent Application Corpus

Masaaki Nagata, Makoto Morishita, Katsuki Chousa, Norihito Yasuda
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai Seika-cho Souraku-gun Kyoto-fu 619-0237 Japan
{masaaki.nagata,makoto.morishita,katsuki.chousa,norihito.yasuda}@ntt.com

Abstract

We constructed JaParaPat (Japanese-English Parallel Patent Application Corpus), a bilingual corpus of more than 300 million Japanese-English sentence pairs from patent applications published in Japan and the United States from 2000 to 2021. We obtained the publication of unexamined patent applications from the Japan Patent Office (JPO) and the United States Patent and Trademark Office (USPTO). We also obtained patent family information from the DOCDB, that is a bibliographic database maintained by the European Patent Office (EPO). We extracted approximately 1.4M Japanese-English document pairs, which are translations of each other based on the patent families, and extracted about 350M sentence pairs from the document pairs using a translation-based sentence alignment method whose initial translation model is bootstrapped from a dictionary-based sentence alignment method. We experimentally improved the accuracy of the patent translations by 20 bleu points by adding more than 300M sentence pairs obtained from patent applications to 22M sentence pairs obtained from the web.

Keywords: Pattent application, Parallel corpus, Japanese-English

1. Introduction

International patent applications are numerous but finite. In this work, we aim to disclose the quantity and the quality of the parallel data obtainable from international patent applications in Japanese and English and the potential translation accuracy using these resources. Since most translation for international patent applications in Japan involves Japanese to English, we focus only on translation from Japanese to English.

Gordon et al. (2021) and Bansal et al. (2022) showed that the accuracy of machine translation improves as the amount of training data or the number of model parameters increases. What makes patent translation different from other machine translation domains is that numerous international patent applications are publicly available after a certain period. However, what we can achieve by exploiting such resources remains unknown.

The history of creating a parallel corpus of Japanese-English patents spans nearly 20 years. Utiyama and Isahara (2007) created a bilingual Japanese-English patent corpus of approximately 2 million sentence pairs for the NTCIR-6 patent retrieval task (Fujii et al., 2007). They applied a bilingual sentence extraction method originally developed for comparable newspaper articles (Utiyama and Isahara, 2003) to patent applications. These bilingual data comprised the first publicly available large-scale Japanese-English patent corpus and were used in the NTCIR-7 patent MT task, which was the first shared task for machine translation between Japanese and English (Fujii et al., 2008).

The JPO-NICT English-Japanese parallel corpus (Japan Patent Office and National Institute of Information and Communications Technology), which has about 350 million Japanese-English patent sentence pairs, was jointly compiled by the Japan Patent Office (JPO) and the National Institute of Information and Communications Technology (NICT) from the publications of unexamined patent applications in the United States and Japan based on patent families. These data, which are available to members of Advanced Language Information Forum (ALAGIN), an organization that resembles LDC, can be used without charge for research and development purposes. The JPO Patent Corpus (Japan Patent Office) has 1M Japanese-English patent sentence pairs and is used in the shared task of patent translation in the Workshop on Asian Translation (WAT), which was first held in 2015.

The JPO-NICT and JPO patent corpora were created around 2015, so they do not reflect the latest contents and technologies. According to Utiyama and Isahara (2007), the JPO-NICT corpus includes JPO and USPTO patents from 1993 but not after 2015. In addition, they were made using a bilingual dictionary-based sentence alignment method (Utiyama and Isahara, 2003). Unfortunately, the quality of dictionary-based alignment (Varga et al., 2005) is generally lower than that of translation-based alignment (Sennrich and Volk, 2010). State-of-the-art sentence alignment technology could improve the quality of Japanese-English patent copora.

We constructed JaParaPat (Japanese-English parallel patent application corpus), which has about 350M sentence pairs from about 1.4M document

LREC-COLING 2024, pages 9452–9462
20-25 May, 2024.

pairs from 2000 to 2021 using translation-based alignment. International patent applications can be filed in one of two ways: the Paris route or the PCT route. To the best of our knowledge, ours is the first attempt to extensively mine parallel patent applications under both routes and align every part of the documents including titles, abstracts, descriptions, and claims[1].

2. Resources

2.1. International Patent Application

There are two ways to obtain a patent in a foreign country: directly filing an application in that country based on the Paris Convention (Paris route) or transferring an international application filed to a patent office based on the Patent Cooperation Treaty (PCT route) to that country.

Under the Paris Convention route, after filing a national application in one country, an application is filed in another country, claiming priority under the Paris Convention within a priority period of one year.

In a PCT application, filing a single PCT application in a single language using a common format to a PCT receiving office secures priority on the filing date in every PCT member country. However, to obtain a patent right in a country, a national phase application must be filed within 30 months of the priority date in that country and an examination of the patent must be undergone following the laws of that country. At that time, the patent application must be translated into the language accepted by that country's patent office.

For example, suppose a Japanese company submits a PCT application written in Japanese to the World Intellectual Property Organization (WIPO). In that case, JPO publishes the Japanese patent application after entry into Japan, and USPTO publishes the English patent application after entry into the United States.

2.2. JPO Patent Data

Since the Japan Patent Office (JPO) provides bulk download service of patent information, [2] we sent the hard drive to the patent office, which returned it with the necessary patent information. If a company uses this system, it must submit a company registry.[3]

In the Japanese Patent Gazette, PCT patent applications are given a different name than ordinary domestic applications. A "published patent application" is an ordinary domestic patent written in Japanese. This is the target of the Paris route searches. A "Japanese translation of PCT international patent application" is a Japanese translation of an international patent application filed with a receiving office other than the JPO for entry into Japan. A "domestic re-publication of PCT international patent application" is an international patent application written in Japanese where JPO is the receiving office.

On December 23, 2021, the JPO abolished the system of publishing domestic re-publication of PCT international patent applications. After this date, PCT applications first filed in Japan in Japanese will only be available if they are granted as a patent after certain amendments, so this study covers the period through 2021.

As shown in the upper part of Figure 1, a JPO XML file represents each patent data by jp-official-gazette element.[4] The kind-of-jp attribute is the gazette type. A is a published patent application, T is a Japanese translation of PCT international patent application, and S is a domestic re-publication of PCT international patent application.

Bibliographic information is found in the bibliographic-data element. For documents whose kind-of-jp attribute is A or T, the publication number is obtained from the publication-reference element and the application number is obtained from the application-reference element. For documents whose kind-of-jp attribute is T, the application number is obtained from the pct-or-regional-filing-data element and the publication number is obtained from the pct-or-regional-publishing-data element.

We extracted the text enclosed by the p tags of the XML elements corresponding to the patent's title, abstract, description, and claim. In other words, for sentence alignment, we excluded the claim numbers, the paragraph numbers, the mathematical expressions, figures, the etc. Since January 2004, Japanese patent applications have been filed in the XML format. Before 2004, they were in the SGML format. We veryfied that data in the SGML format have the same extraction targets as in the XML format.

[1]We will make a part of JaParaPat (years 2016-2020, about 110M sentence pairs) publicly available for research purposes after our paper is published.

[2]https://www.jpo.go.jp/system/laws/sesaku/data/download.html

[3]Although JPO's web page do not mention license conditions, we confirmed with the organization that we can use these data for the research and the development of machine translation.

[4]https://www.jpo.go.jp/system/laws/koho/shiyo/kouhou_siyou_vol4-7.html

```xml
<?xml version="1.0" encoding="EUC-JP"?>
<?xml-stylesheet type="text/xsl" href="../../../../../XSL/gat-a.xsl"?>
<!DOCTYPE jp-official-gazette PUBLIC "-//JPO//DTD PUBLISHED PATENT/UTILITY MODEL
 APPLICATION 1.0//EN" "../../../../../DTD/gat-a.dtd">
<jp-official-gazette kind-of-jp="A" kind-of-st16="A" lang="ja" dtd-version="1.0"
 country="JP" xmlns:jp="http://www.jpo.go.jp"><bibliographic-data lang="ja" coun
try="JP">
    <publication-reference>
      <document-id>
        <country>JP</country>
        <doc-number>2021093912</doc-number>
        <kind>公開特許公報(A)</kind>
        <date>20210624</date>
      </document-id>
    </publication-reference>
    <application-reference>
      <document-id>
        <doc-number>2018058673</doc-number>
        <date>20180326</date>
      </document-id>
    </application-reference>
    <invention-title>検体中に含まれる菌種を特定する方法</invention-title>
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE us-patent-application SYSTEM "us-patent-application-v46-2022-02-17.dtd
" [ ]>
<us-patent-application lang="EN" dtd-version="v4.6 2022-02-17" file="US202202956
84A1-20220922.XML" status="PRODUCTION" id="us-patent-application" country="US" d
ate-produced="20220907" date-publ="20220922">
<us-bibliographic-data-application lang="EN" country="US">
<publication-reference>
<document-id>
<country>US</country>
<doc-number>20220295684</doc-number>
<kind>A1</kind>
<date>20220922</date>
</document-id>
</publication-reference>
<application-reference appl-type="utility">
<document-id>
<country>US</country>
<doc-number>17619810</doc-number>
<date>20200617</date>
</document-id>
</application-reference>
```


Figure 1: Example of JPO and USPTO XML files

2.3. USPTO Patent Data

The United States Patent and Trademark Office (USPTO) provides patent application full text data. [5] We can obtain the documentation and the DTD from USPTO's web page.[6] USPTO provides patent application full text data from March 15, 2001. Since corresponding patent applications may have been published in Japan one year before they were published in the U.S., this study covers the period from 2000.

As shown in the lower part of Figure 1, a USPTO XML file represents each patent by us-patent-application element. Bibliographic information is in the us-bibliographic-data-application element. The application number is obtained from the application-reference element, and the publication number is obtained from the publication-reference element.

If a pct-or-regional-filing-data element exists and its doc-number attribute begins with PCT, such as "PCT/JP2005/003817," we consider it a PCT application, and the value of the doc-number attribute is its application number. The USPTO's PCT patent application does not have the same distinction as that between T and S in the JPO's kind-of-jp attribute.

2.4. EPO DOCDB

The European Patent Office (EPO) provides (for a fee) worldwide bibliographic data of patents (DOCDB). We can obtain a sample of DOCDB[7] and its manual [8] from its web site.

We obtained DOCDB, as of April 2022, to get information on patent families. A patent family is a set of patents obtained in various countries to

[5] https://developer.uspto.gov/product/patent-application-full-text-dataxml

[6] https://www.uspto.gov/learning-and-resources/xml-resources

[7] https://www.epo.org/searching-for-patents/data/bulk-data-sets/docdb.html

[8] https://www.epo.org/searching-for-patents/data/bulk-data-sets/manuals.html

```
[
  "country": "WO",
  "kind": "A1",
  "doc-id": "550103527",
  "doc-number": "2021085030",
  "family-id": "75715054",
  "date-publ": "20210506",
  "is-representative": "YES",
  "originating-office": "EP",
  "title": "DRIVING ASSISTANCE SYSTEM",
  "publication-reference": {
    "country": "WO",
    "doc_number": "2021085030"
  },
  "application-reference": {
    "doc_id": "550103526",
    "country": "JP",
    "doc_number": "2020037511",
    "kind": "W"
  },
  "priority-claims": [
    {
      "doc-id": null,
      "country": "US",
      "doc_number": "201962927868",
      "kind": "P",
      "date": "20191030"
    }
  ],
  "patent-family": []
],
```


Figure 2: Example of information extracted from exch:exchange-document element in DOCDB

protect a single invention. We obtained a patent family by analyzing the priority claim data in the DOCDB.

A DOCDB XML file represents each patent by exch:exchange-document element. The priority-claims information is aggregated in the exch:priority-claims element under the exch:bibliographic-data element. Figure 2 is an example of information extracted from exch:exchange-document element in a DOCDB XML file. We extracted the country, doc-number, kind, and date attributes of the document-id element, which is the subject of the priority claim. Kind-code A is an ordinary patent application, and W is a PCT application.

3. Methodology

3.1. Document Alignment

We mapped the patent applications published by the JPO and USPTO based on the patent families obtained from the EPO's DOCDB. The original data are all in XML, and we implemented the document alignment procedure described below using the xml.etree.ElementTree module in the python standard library.

We considered pairs of Japanese and English patent applications in the same patent family to be translations of each other. If there are more than one such pairs, we selected the oldest document pair because a set of documents claiming priority for the same document is almost always a modified version of the initial application.

The search method for a bilingual document pair differs slightly between the Paris route and the PCT routes. The primary example of the Paris route is where one application claims priority based on another. A US patent that claims priority based on one filed in Japan is a patent in DOCDB where the country attribute of the exchange-document element is US and the country attribute and kind attribute in the priority-claims element are JP and A, respectively. The same is true for a Japanese patent that claims priority based on one filed in the US. In this paper, we refer to the former as 'jp-us' and the latter as 'us-jp' based on the order in which the patents were filed in the countries.

We extracted a pair of Japanese and U.S. patent applications that claims priority based on a shared third patent application, such as a patent that is first filed in China and then filed in Japan and the U.S. For these cases, we first listed a pair of the document-id in the exchange-document element and the document-id in the priority-claim element, for all Japanese and U.S. patent applications. We then extracted JP-US patent application pairs with the same document-id in the priority-claim element. In this paper, we refer to such pairs as 'jp-x-us' where x indicates that a shared third patent application exists.

For the PCT route, we first extracted from the DOCDB applications where the kind attribute of the application-reference element is W. We extracted applications from the JPO where the kind-code attribute is S or T and the doc-number starts with WO. We extracted applications from the USPTO where the pct-or-regional-filing-data starts with PCT. If the application number obtained from the JPO data and the application number obtained from the USPTO data are the same and exist in the DOCDB, we consider the Japanese and the U.S. patent applications to be translations of each other. In this paper, we refer to all PCT applications as 'pct'.

3.2. Sentence Alignment

We used two methods for sentence alignment: one based on bilingual dictionaries (Utiyama and Isahara, 2003) and another based on machine translation (Sennrich and Volk, 2010). We first obtained a bilingual patent data using a dictionary-based sentence alignment method and trained a translation model from the bilingual patent data and JParaCrawl (Morishita et al., 2022), a publicly available large-scale Japanese-English parallel corpus collected from the web. We then obtained the final bilingual patent data using a translation-based sentence alignment method.

We divided the Japanese and U.S. patent applications into titles, abstracts, descriptions, and claims and aligned them separately. We used split-sentences.perl in Moses for sentence segmentation in both Japanese and English. [9]

For our dictionary-based sentence alignment, we used our implementation of Utiyama and Isahara (2003)'s method. As for bilingual dictionary, we used a Japanese-English dictionary of EDR with 1,690,174 entries (Japan Electronic DIctionary Research Institute, Ltd.). We used mecab-unidic[10] for Japanese word segmentation and TreeTagger[11] for English tokenization.

For translation-based sentence alignment, we used Bleualign[12]. We used fairseq (Ott et al., 2019) for machine translation.

4. JaParaPat Overview

4.1. Data Statistics

Table 1 shows the number of annually collected document and sentence pairs from 2000 to 2001. In this table, the numbers are divided into jp-us, jp-x-us, us-jp, and pct, as described in Section 3.1. Here, the years are based on the publication year of the Japanese patent applications.

The parallel corpus has about 350M sentence pairs from about 1.4M document pairs. Since the USPTO U.S. patent data are only available after 2001, no Japanese patent applicationss published in 2000 have any available U.S. patent applications as priority claims. Since we used the DOCDB as of April 2022, the patent family are incomplete on the applications published in Japan in 2021. Thus, scant parallel data exist for 2021.

The ratio of Paris routes to PCT routes in the parallel corpus is almost one-to-one. The former route has more document pais, but the latter route has more sentence pairs because document pairs in the Paris route are not necessarily translations of each other, while document pairs in the PCT route must be translations of each other. In general, we extracted 60-70% of the sentences as parallel sentence pairs from the Japanese and English document pairs. Within the Paris route, The amount of bilingual data for us-jp is the largest, followed by jp-us and jp-x-us.

4.2. Data Format

Figure 3 shows an example of Japanese and English text files for a patent document pair. We first assigned a pair of publication numbers in Japan and the U.S. as an ID for a parallel document pair, such as JP2021000998-US20210139186. We divided Japanese and U.S. patent documents into four parts: title, abstract, description, and claim, separated each part into paragraphs and sentences, and finally assigned a concatenation of a document pair ID, a part, a paragraph number, a sentence number within a paragraph, and a sentence number within a document as an ID to a sentence.

The leftmost screenshot in Figure 4 shows an example of a sentence alignment file for a patent document pair. The first column represents the sentence number within a Japanese document, and the second column represents the sentence number within an English document. Multiple numbers in one column represent a many-to-many alignment. This configuration allows us to create a claim-specific translation model by extracting only the sentence pairs in the claim, or a context-aware translation model by extracting consecutive sentence pairs in the same paragraph.

The middle and rightmost screenshots in Figure 4 shows examples of International Patent Classification (IPC) data for each document pairs. This information allows us to create a translation model dedicated to a specific field.

5. Experiments

5.1. Training and Test Data

To confirm the quality of JaParaPat, we conducted translation experiments from Japanese to English. Table 2 shows the number of document pairs, sentence pairs, and the number of words on the English side of the training data for the translation model. We used the sentence pairs from 2000 to the first half of 2021 to train the translation models.

Table 3 shows the number of sentences and words on the English side of the test data. We randomly sampled 1,000 sentences for the test data and 2,000 sentences for the validation data from the second half of 2021 in the Paris and PCT routes, respectively. Note that while these Paris and PCT test sets cover a wide range of topics, they are not guaranteed to be parallel sentence pairs because they are automatically extracted and sampled.

We also used as test data the in-house Japanese PCT patent applications published or to be published in 2022 or later and their translations into English by two translation companies specializing in patent translation. The target domain is information and communication technology (ICT) and includes a wide range of content from hardware to software. Preliminary studies revealed that the scores of automated evaluations varied by trans-

[9] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/ems/support/split-sentences.perl

[10] https://taku910.github.io/mecab/

[11] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[12] https://github.com/rsennrich/Bleualign

	Sentence pairs				Document pairs			
	jp-us	jp-x-us	us-jp	pct	jp-us	jp-x-us	us-jp	pct
2000	804,586	116,806		92,242	4,189	865		402
2001	1,936,229	423,355	842,701	122,205	11,223	3,249	5,608	550
2002	2,599,128	1,161,071	3,181,974	51,214	14,385	8,521	18,941	200
2003	2,216,059	1,944,235	4,083,604	1,975,669	11,755	12,506	22,385	7,743
2004	2,719,911	860,287	3,848,196	4,319,575	16,126	7,542	23,324	18,978
2005	2,352,235	994,049	5,024,330	4,977,803	12,973	8,193	28,089	20,647
2006	2,297,878	1,131,340	5,770,905	4,513,947	12,239	8,810	30,832	18,469
2007	2,513,900	1,081,103	5,883,197	5,050,197	13,124	8,147	30,481	20,444
2008	2,535,483	921,678	5,752,965	8,264,349	12,956	6,715	29,165	31,506
2009	1,813,767	861,456	6,259,067	8,227,809	9,180	6,049	31,303	31,304
2010	1,559,327	821,388	6,310,667	8,178,496	7,381	5,169	29,025	29,196
2011	1,869,428	957,781	6,739,639	6,497,215	8,341	5,789	28,899	22,932
2012	1,990,833	945,927	7,252,931	7,781,432	8,868	5,560	30,065	27,381
2013	2,363,076	1,012,462	6,598,196	10,278,504	10,050	6,021	28,101	35,850
2014	2,144,452	1,116,288	6,651,888	8,055,146	9,168	6,088	26,716	27,326
2015	2,506,286	1,030,098	6,754,694	9,391,589	10,314	5,229	26,087	31,380
2016	2,494,488	1,017,181	5,746,295	9,313,031	10,233	4,988	22,317	29,196
2017	4,861,052	1,017,358	3,624,756	16,251,900	19,876	5,045	14,467	51,791
2018	3,284,674	918,138	5,153,238	11,696,010	12,625	4,369	19,239	35,822
2019	3,227,271	1,066,833	6,107,334	12,483,342	12,388	5,251	23,685	36,961
2020	3,740,996	1,093,506	4,251,027	11,962,022	13,306	4,781	15,032	34,006
2021	1,043,944	849,489	4,838,957	11,275,167	3,656	3,818	16,928	30,884
sum	52,875,003	21,341,829	110,676,561	160,758,864	244,356	132,705	500,689	542,968
	345,652,257				1,420,718			

Table 1: Number of parallel sentence and document pairs collected annually from 2000 to 2021

route	documents	sentences	words
Paris	866,931	181,907,843	7,378,214,793
PCT	527,068	154,860,596	6,180,045,629
Paris+PCT	1,393,999	336,768,439	13,558,260,422

Table 2: Number of document pairs, sentence pairs, and words on English side in the training data

Test data	#sentences	#words
Paris SH2021	1,000	37,990
PCT SH2021	1,000	38,676
In-house test1	1,002	33,405
In-house test2	988	26,945
ASPEC test	1,812	39,573

Table 3: Number of sentences and words on English side in the test sets

lation companies, not by content, so we created a test set for each translation company.

We also used test sentences from the Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016) as publicly available out-of-domain test data. There are no publicly available in-domain (patent) test data suitable for the quality assessment of our parallel corpus. Since our training data covers from 2000 to the first half of 2021, the test data should be Japanese patent applications published in the second half of 2021 or later. However, the JPO Patent Corpus test set used in the patent translation shared task of WAT-2023 was made from patent documents published in 2019-2020, which was likely to be included in our training data.

5.2. Training Conditions

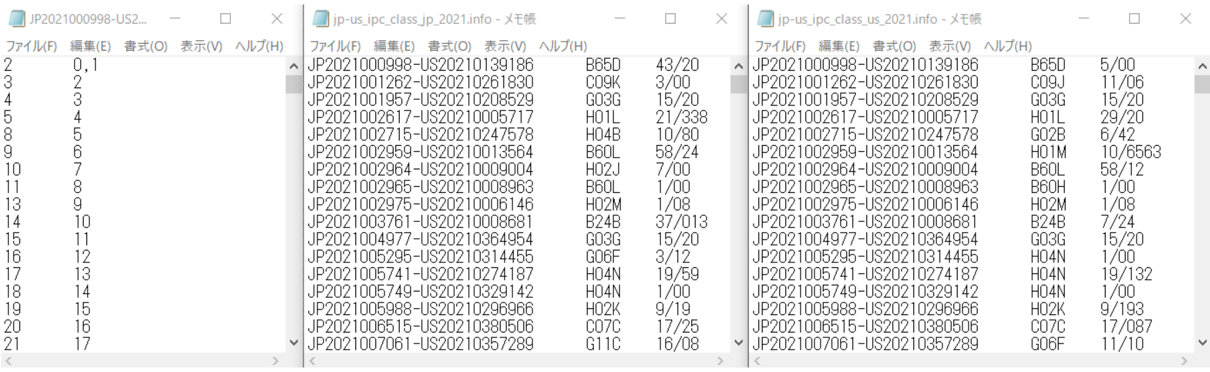
architecture	transformer_wmt_en_de_big
enc-dec layers	6
optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98$)
learning rate schedule	inverse square root decay
warmup steps	4,000
max learning rate	0.001
dropout	0.3
gradient clip	0.1
batch size	1M tokens
max number of updates	60K steps
validate interval updates	1K steps
patience	5

Table 4: List of hyperparameters for the Transformer

```
JP2021000998-US20210139186.ja - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
JP2021000998-US20210139186_title_0000_0_0	収納ケース
JP2021000998-US20210139186_abstract_0000_0_1	【課題】剛性を高くする。
JP2021000998-US20210139186_abstract_0000_1_2	【解決手段】収納ケース１０は、前側及び上側へ開放された直方体箱状のケース
JP2021000998-US20210139186_abstract_0000_2_3	これにより、枠部材５０によって、収納ケース１０の前端部における剛性を高く
JP2021000998-US20210139186_abstract_0000_3_4	また、開閉パネル６０が、枠部材５０に回転可能に組付けられている。
JP2021000998-US20210139186_abstract_0000_4_5	そして、開閉パネル６０が、閉位置から前側へ回転されることで回転位置に配置
JP2021000998-US20210139186_abstract_0000_5_6	これにより、収納物を収納ケース１０から出し入れするときの使用者に対する利
JP2021000998-US20210139186_abstract_0000_6_7	【選択図】図１５
JP2021000998-US20210139186_description_0000_0_8	本発明は、収納ケースに関する。
JP2021000998-US20210139186_description_0001_0_9	下記特許文献１に記載の収納ケースは、上側及び前側へ開放されたボックス本体
JP2021000998-US20210139186_description_0001_1_10	開閉蓋部は、前面開口部を覆う閉塞位置と、天板部とボックス本体の側
JP2021000998-US20210139186_description_0001_2_11	そして、開閉蓋部の開放位置では、開閉蓋部が、ボックス本体の内部に
JP2021000998-US20210139186_description_0002_0_12	特開２０１６－９４２３９号公報
JP2021000998-US20210139186_description_0003_0_13	しかしながら、上記収納ケースでは、天板部をボックス本体に組付けた
JP2021000998-US20210139186_description_0003_1_14	このため、収納ケースの剛性を高くするという点において改善の余地が
JP2021000998-US20210139186_description_0004_0_15	本発明は、上記事実を考慮して、剛性を高くすることができる収納ケー
JP2021000998-US20210139186_description_0005_0_16	本発明の１又はそれ以上の実施形態は、前側及び上側へ開放された直方
JP2021000998-US20210139186_description_0006_0_17	本発明の１又はそれ以上の実施形態は、前記開閉パネルの上端部には、
JP2021000998-US20210139186_description_0007_0_18	本発明の１又はそれ以上の実施形態は、前記枠部材には、前記収容位置
JP2021000998-US20210139186_description_0008_0_19	本発明の１又はそれ以上の実施形態は、前後方向における開閉パネルの
JP2021000998-US20210139186_description_0009_0_20	本発明の１又はそれ以上の実施形態によれば、剛性を高くすることがで
JP2021000998-US20210139186_description_0010_0_21	本実施の形態に係る収納ケースを示す斜視図である。
JP2021000998-US20210139186_description_0010_1_22	図１に示される収納ケースのスタッキング状態を示す斜視図である。
```

```
JP2021000998-US20210139186.en - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
JP2021000998-US20210139186_title_0000_0_0	STORAGE BOX
JP2021000998-US20210139186_abstract_0000_0_1	A storage box is configured to include a rectangular parallelepiped box-s
JP2021000998-US20210139186_abstract_0000_1_2	As a result, the rigidity of the front end part of the storage box can be
JP2021000998-US20210139186_abstract_0000_2_3	In addition, the opening-closing panel is rotatably assembled to the fram
JP2021000998-US20210139186_abstract_0000_3_4	The opening-closing panel is disposed at the rotation position by being r
JP2021000998-US20210139186_description_0000_0_5	The present disclosure relates to a storage box.
JP2021000998-US20210139186_description_0001_0_6	A storage box disclosed in JP-A-2016-94239 below is configured to include
JP2021000998-US20210139186_description_0001_1_7	The lid opening-closing portion is configured to be movable between a clo
JP2021000998-US20210139186_description_0001_2_8	At the open position of the lid opening-closing portion, since the lid op
JP2021000998-US20210139186_description_0002_0_9	However, in the storage box described above, when the top plate portion i
JP2021000998-US20210139186_description_0002_1_10	Therefore, there is a room for improvement in terms of increasing
JP2021000998-US20210139186_description_0003_0_11	An object of the present disclosure is to provide a storage box i
JP2021000998-US20210139186_description_0004_0_12	According to one or more embodiments of the present disclosure, a
JP2021000998-US20210139186_description_0005_0_13	In the storage box according to one or more embodiments of the pr
JP2021000998-US20210139186_description_0006_0_14	In the storage box according to one or more embodiments of the pr
JP2021000998-US20210139186_description_0007_0_15	In the storage box according to one or more embodiments of the pr
JP2021000998-US20210139186_description_0008_0_16	According to one or more embodiments of the present disclosure, t
JP2021000998-US20210139186_description_0009_0_17	FIG.1 is a perspective view illustrating a storage box according
JP2021000998-US20210139186_description_0010_0_18	FIG.2 is a perspective view illustrating a stacking state of the
JP2021000998-US20210139186_description_0011_0_19	FIG.3 is a side sectional view illustrating a state in which the
JP2021000998-US20210139186_description_0012_0_20	FIG.4 is a side sectional view (a sectional view taken along line
JP2021000998-US20210139186_description_0013_0_21	FIG.5 is a sectional view (a sectional view taken along line V-V
JP2021000998-US20210139186_description_0014_0_22	FIG.6 is a perspective view of the case main body illustrated in
```

Figure 3: Example of Japanese and English text files for a patent document pair



```
JP2021000998-US2...
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
2	0,1
3	2
4	3
5	4
8	5
9	6
10	7
11	8
13	9
14	10
15	11
16	12
17	13
18	14
19	15
20	16
21	17
```

```
jp-us_ipc_class_jp_2021.info - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
JP2021000998-US20210139186	B65D	43/20
JP2021001262-US20210261830	C09K	3/00
JP2021001957-US20210208529	G03G	15/20
JP2021002617-US20210005717	H01L	21/338
JP2021002715-US20210247578	H04B	10/80
JP2021002959-US20210013564	B60L	58/24
JP2021002964-US20210009004	H02J	7/00
JP2021002965-US20210008963	B60L	1/00
JP2021002975-US20210006146	H02M	1/08
JP2021003761-US20210008681	B24B	37/013
JP2021004977-US20210364954	G03G	15/20
JP2021005295-US20210314455	G06F	3/12
JP2021005741-US20210274187	H04N	19/59
JP2021005749-US20210329142	H04N	1/00
JP2021005988-US20210296966	H02K	9/19
JP2021006515-US20210380506	C07C	17/25
JP2021007061-US20210357289	G11C	16/08
```

```
jp-us_ipc_class_us_2021.info - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
JP2021000998-US20210139186	B65D	5/00
JP2021001262-US20210261830	C09J	11/06
JP2021001957-US20210208529	G03G	15/20
JP2021002617-US20210005717	H01L	29/20
JP2021002715-US20210247578	G02B	6/42
JP2021002959-US20210013564	H01M	10/6563
JP2021002964-US20210009004	B60L	58/12
JP2021002965-US20210008963	B60H	1/00
JP2021002975-US20210006146	H02M	1/08
JP2021003761-US20210008681	B24B	7/24
JP2021004977-US20210364954	G03G	15/20
JP2021005295-US20210314455	H04N	1/00
JP2021005741-US20210274187	H04N	19/132
JP2021005749-US20210329142	H04N	1/00
JP2021005988-US20210296966	H02K	9/193
JP2021006515-US20210380506	C07C	17/087
JP2021007061-US20210357289	G06F	11/10
```

Figure 4: Example of sentence alignment file for a patent document pair and IPC data for patent document pairs

We used fairseq (Ott et al., 2019) for machine translation. The translation model is Transformer big (Vaswani et al., 2017). Table4 shows the hyperparameters of the Transformer. The translation models in this paper were all trained under this condition. We used sentencepiece (Kudo and Richardson, 2018) for tokenization. We randomly sampled 7M sentence pairs from the patent corpus and 3M sentence pairs from JParaCrawl to train the sentencepiece model. The vocabulary size was 32K for both Japanese and English. We set the character_coverage to 0.9995 and the byte_fallback to true. We used both sacreBLEU (Papineni et al., 2002; Post, 2018) and COMET (Rei et al., 2020) for evaluation, but we mainly used BLEU because choosing the appropriate technical terms is essential in patent translation.

5.3. Comparison of Sentence Alignment Methods

First, we examined the accuracy of the translation model used in our translation-based sentence alignment method. We collected about 34M sentence pairs (2000-2013Paris_dict) from document pairs in the Paris route from 2000 to 2013 using a dictionary-based sentence alignment method. We then created a translation model trained on these 34M patent sentence pairs and JParaCrawl (2000-2013Paris_dict+JParaCrawl). Using this translation model for translation-based sentence alignment, we collected about 43M sentence pairs (2000-2013Paris_trans) from the same document pairs used for dictionary-based sentence alignment.

Table5 shows that the translation accuracy (BLEU) improved when we combined the sentence pairs from the patent applications and the web. When we use translation-based sentence alignment, we collected more sentence pairs (34M to 43M) with higher quality (62.6/51.5 to 63.4/53.0) than dictionary-based sentence alignment. Recent research shows that translation-based sentence alignment method can obtain better and more bilingual sentence pairs than dictionary-based method (Bañón et al., 2020; Morishita et al., 2022) and we confirmed this finding in our experiment.

Test1 and test2 differed in sacreBLEU by 10 points in the models trained on the patent corpus. Since both translation companies manually post-edited the output of their patent translation systems, we assume that the differences in the machine translation and post-editing methods significantly impacted the automatic evaluation measurements. The results indicate that post-editing bias may be a problem in the future for parallel corpora collected from patent applications because more and more patent translation companies are adopting machine translation post-editing.

5.4. Japanese-to-English Translation Accuracy

Table 6 shows the translation accuracy of the model trained from the collected patent sentence pairs. Compared to JParaCrawl, JaParaPat improved the patent translation accuracy by 20 bleu points. Comparing the Paris route and the PCT routes, although the amount of data is almost identical (around 150M), the Paris route has generally higher translation accuracy. We assume this result is because the Paris route contains a greater variety of patent applications since the PCT route is mainly used by large companies.

Training the translation model from more than 300M patent bilinguals from both the Paris and PCT routes improved translation accuracy, although the improvement is moderate and unstable. However, when we added 22M web-crawled sentence pairs of JParaCrawl to 337M patent sentence pairs of JaParaPat, the translation accuracy of test2 and ASPEC increased, suggesting that the patent sentence pairs lack diversity. We observed that the perplexity of the patent texts is low compared to that of web texts. Adding web text makes the patent translation model more robust than increasing the amount of patent text.

6. Related Works

6.1. Patent Parallel Corpus

With the increasing popularity of the PCT international patent applications and such new technologies as sentence alignment using neural machine translation models, a different approach has recently emerged for creating a parallel patent corpus. In 2011, World Intellectual Property Organization (WIPO) created the Corpus Of Parallel Patent Applications (COPPA) from the titles and abstracts of PCT applications. COPPA V2.0 (Junczys-Dowmunt et al., 2016) consists of eight language pairs, mainly English, and has about 1 million sentence pairs of Japanese-English data. ParaPat (Soares et al., 2020) is a bilingual data set of 22 language pairs created from patent abstracts in Google Patents, with 17M sentence pairs of Japanese-English data. COPPA and ParaPat use Hunalign (Varga et al., 2005), a dictionary-based sentence alignment tool.

EuroPat (Heafield et al., 2022) is a parallel patent corpus of six European language as well as English collected from USPTO and EPO. It extracts sentence pairs from granted patents with an emphasis on quality. It uses the API provided by the EPO to obtain patent families for document alignment and translates non-English documents into English for sentence alignment with Bleualign-cpp[13], a translation-based sentence alignment tool developed in the ParaCrawl Project (Bañón et al., 2020).

Our approach resembles EuroPat, although we used unexamined patent applications rather than granted patents and made alignments between

[13] https://github.com/bitextor/bleualign-cpp

training data	test1	test2	pairs	updates
2000-2013Paris_dict	62.6	51.5	34M	17K
2000-2013Paris_dict+JParaCrawl	63.6	54.0	56M	26K
2000-2013Paris_trans	63.4	53.0	43M	16K

Table 5: Comparison of Sentence Alignment Methods

training data	Paris		PCT		test1		test2		ASPEC		pairs	updates
	bleu	comet	bleu	comet	bleu	comet	bleu	comet	bleu	comet		
JParaCrawl(JPC)	31.9	0.817	35.6	0.827	36.2	0.838	35.8	0.826	20.6	**0.828**	22M	20K
Paris	**55.6**	**0.867**	56.5	**0.877**	66.8	**0.881**	53.2	0.820	20.5	0.823	182M	44K
PCT	52.7	0.857	**57.3**	0.873	64.6	0.866	51.6	0.811	20.6	0.820	155M	53K
Paris+PCT	55.5	0.864	55.7	0.872	67.0	0.876	46.0	0.820	20.8	0.821	337M	57K
JPC+Paris+PCT	54.7	0.863	56.0	0.872	**67.7**	0.880	**55.5**	**0.846**	**21.3**	0.827	359M	42K

Table 6: Comparison of translation accuracies with respec to the size of training data

Japanese and English rather than among European languages.

6.2. Japanese-English Parallel Corpus

In areas other than patents, ASPEC (Nakazawa et al., 2016) is one of the first publicly available Japanese-English parallel corpora. It is comprised of English summaries attached to Japanese scientific and technical papers. Its domain is close to patents, but it only has 3 million sentence pairs. ASPEC has been used in a shared task of WAT since 2014 (Nakazawa et al., 2014).

JParaCrawl (Morishita et al., 2022) is one of the largest publicly available Japanese-English parallel corpora. It is a web-crawled corpus that contains a wide variety of domains. JParaCrawl has been used in news translation task and general machine translation task in WMT since 2020 (Barrault et al., 2020).

Although the JPO-NICT corpus is one of the largest publicly available Japanese-English parallel patent corpora, its construction is unknown since it has not been published as a technical paper. Assuming that this corpus was made from a procedure similar to the NICIR-7 PATMT (Utiyama and Isahara, 2007), it identifies Japanese patents by the priority number listed in the U.S. patents. Thus, this corpus only covers the jp-us of the Paris route in our term. It used dictionary-based sentence alignment, while we used sentence-based alignment.

The newly created JaParaPat is one of the largest and highest-quality Japanese-English patent parallel corpora. It will serve as the foundation for future machine translation research in the science and technology field.

6.3. Sentence Alignment

Sentence alignment can be classified into three categories: a bilingual dictionary-based method (Utiyama and Isahara, 2003; Varga et al., 2005) such as hunalign, a machine translation-based method (Sennrich and Volk, 2010) such as Bleualign, or a multilingual sentence embedding-based method (Thompson and Koehn, 2019; Chousa et al., 2020) such as Vecalign.

Although the sentence embedding-based method is the most accurate approach, it is unfortunately also the most computationally expensive. Since we must process a large amount of data in this work, we used a translation-based method to balance speed and accuracy.

7. Conclusion

We extracted patent sentence pairs as exhaustively as possible from Japanese and U.S. patent applications from 2000-2021 and constructed a parallel patent corpus of more than 300M sentence pairs.

By training a translation model on the parallel patent corpus, we improved the patent translation accuracy by about 20 bleu points compared to JParaCrawl by using 22M sentence pairs collected from the web. We collected more and better sentence pairs by using a translation-based sentence alignment method compared to a dictionary-based sentence alignment method.

Future work includes increasing the number of parameters in the translation model and designing a filter to remove noise in the parallel corpus to improve translation accuracy with reference to the study of data scaling laws (Gordon et al., 2021; Bansal et al., 2022).

8. Bibliographical References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in nmt: The effect of noise and architecture. In *Proceedings of the 39th International Conference on Machine Learning*, pages 1466–1482.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. SpanAlign: Sentence alignment method based on cross-language span prediction and ILP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007. Overview of the patent retrieval task at the ntcir-6 workshop. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 359–365.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2008. Overview of the patent translation task at the ntcir-7 workshop. In *Proceedings of NTCIR-7 Workshop Meeting*, pages 389–400.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenneth Heafield, Elaine Farrow, Jelmer van der Linde, Gema Ramírez-Sánchez, and Dion Wiggins. 2022. The EuroPat corpus: A parallel corpus of European patent data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 732–740, Marseille, France. European Language Resources Association.

Marcin Junczys-Dowmunt, Bruno Pouliquen, and Christophe MazencAndrew. 2016. Coppa v2.0: Corpus of parallel patent applications building large parallel corpora with gnu make. In *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora Workshop Programme*, pages 15–19.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the 1st workshop on Asian translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*, pages 1–19, Tokyo, Japan. Workshop on Asian Translation.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Felipe Soares, Mark Stevenson, Diego Bartolome, and Anna Zaretskaya. 2020. ParaPat: The multi-million sentences parallel corpus of patents abstracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3769–3774, Marseille, France. European Language Resources Association.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan. Association for Computational Linguistics.

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP-2005*, pages 590–596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the NeurIPS 2017*, pages 5998–6008.

9. Language Resource References

Japan Electronic DIctionary Research Institute, Ltd. *The EDR Electronic Dictionary*. National Institute of Information and Communications Technology, ISLRN https://www2.nict.go.jp/ipp/EDR/ENG/indexTop.html.

Japan Patent Office. *JPO Patent Corpus*. Workshop on Asian Translation, ISLRN https://lotus.kuee.kyoto-u.ac.jp/WAT/patent/.

Japan Patent Office and National Institute of Information and Communications Technology. *JPO-NICT English-Japanese Parallel Corpus*. Advanced Language Information Forum, ISLRN https://alaginrc.nict.go.jp/jpo-outline.html.