JaParaPat:
大規模日英特許対訳コーパス

2024/11/7(木) 第8回特許情報シンポジウム

NTT コミュニケーション科学基礎研究所 永田 昌明

はじめに

大規模日英特許対訳コーパス JaParaPat (Japanese-English Parallel Patent Application Corpus)

- 日本と米国の特許出願データ (2000-2021) から約100万文書対の対訳文書を収集し、約3億文対の対訳コーパスを作成
- パリルートとPCTルートの両方から対訳文書を収集
- ブートストラップ: 対訳辞書を用いた文対応 → 対訳コーパス → 機械翻訳モデルを訓練 → 機械翻訳を用いた文対応 → 対訳コーパス
- 約2000万文対のWeb対訳 (JParaCrawl v3.0) に比べ、特許翻訳の精度が 約 20 BLEU ポイント向上

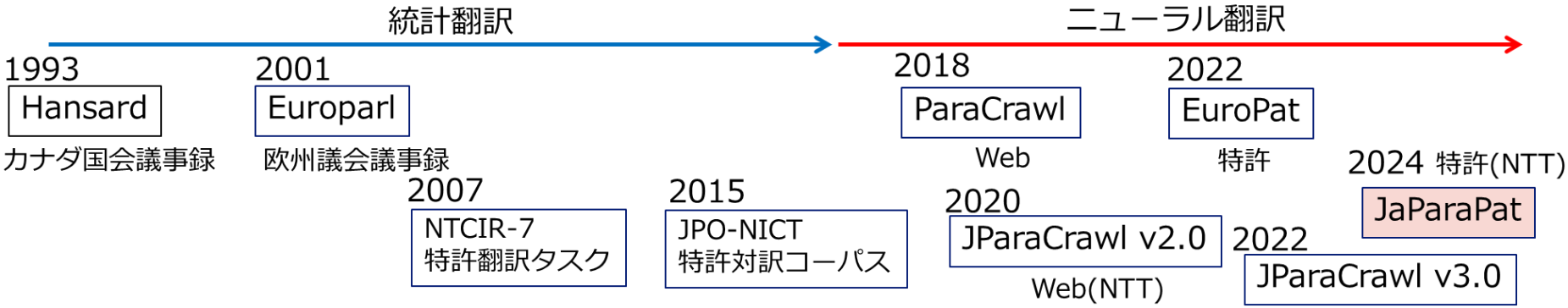
背景1: 日英の特許対訳コーパスの最新化

欧米では特許対訳コーパスの整備が進んでいる

- EuroPat (2022): 6つの欧州言語 (独仏西など)と英語の対訳コーパス。ParaCrawlプロジェクトの技術を応用

日本の特許対訳コーパスは内容も技術も古い

- NTCIR-7 特許翻訳タスク (2008): 約200万文対の日英対訳コーパス
- JPO・NICT英日対訳コーパス (2015): 約3.5億文対 (ALAGINから配布)



背景2: 特許翻訳作業の効率化

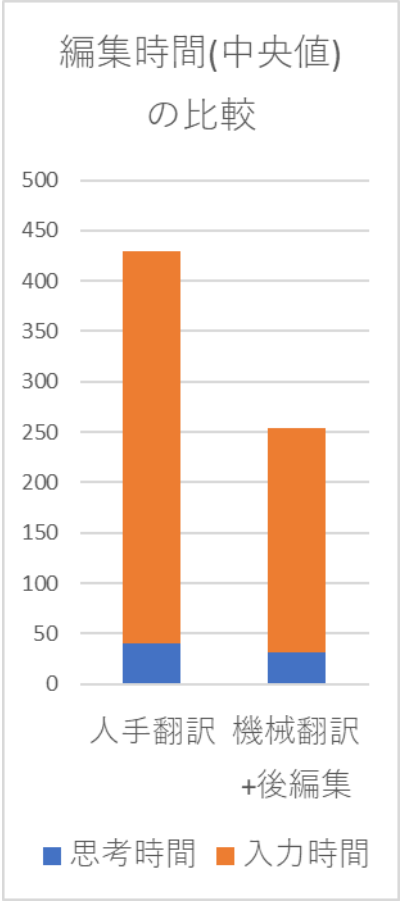
予備実験では機械翻訳の導入で作業時間を約60%に削減

- 約4000文字の原稿 4セットを、4名の翻訳者に順番を変えて提示し、以下の時間(分)を測定
  - 編集時間: 思考時間+入力時間
  - 思考時間: 翻訳開始から最初の編集操作まで
  - 入力時間: 最初の編集操作から次の文の翻訳開始まで

本日の話題

特許翻訳作業の効率化のための研究開発

- 日英特許対訳コーパス [Nagata+, LREC-COLING-2024]
- 語彙制約付き機械翻訳 [Chousa and Morishita, WAT-2021]
- 編集モデルに基づく後編集支援 [Deguchi+,EAMT-2024]



アプローチ: (J)ParaCrawl の技術を特許に適用

JParaCrawl [Morishita+,LREC-2020/2022]

- Webから収集した日英対訳コーパス
  - 4.8M (v1), 8.8M (v2), 21.4M (v3), 44.2M (v4)
- 基本的にはParaCrawl のコーパス構築技術を日本語化

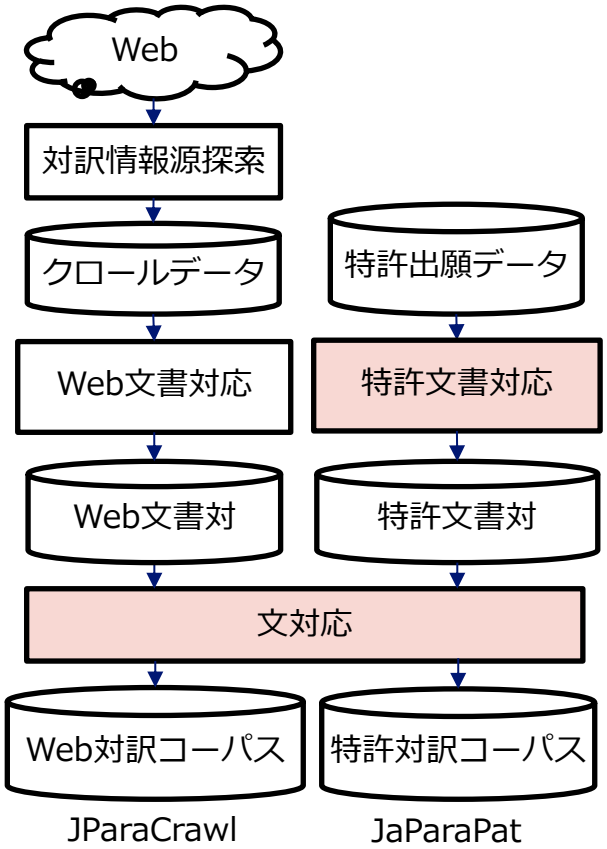
特許文書対応

- パリルートとPCTルートに対応した独自手法

文対応

- 対訳辞書に基づく手法 hunalign [Varga+, RANLP-2005] からブートストラップして、機械翻訳に基づく手法 Bleualign [Sennrich+, AMTA-2010] へ

hunalignの精度 > Bleualignの精度



国際特許出願制度

パリ ルート (パリ条約: Paris Convention)

- 最初にある国へ出願し、優先権主張して一年以内に別の国へ出願
  - › 本研究では最初に出願した国で分類: 日本 (jp-us), 米国 (us-jp), その他 (jp-x-us)

PCT ルート (特許協力条約: Patent Cooperation Treaty)

- PCT受理官庁(receiving office)へ単一の言語と形式で出願し、30カ月以内に国内移行(national phase application)

パテントファミリー

- 同一の発明を保護するために複数の国へ出願された特許の集合

一般に公開されている特許データ

米国特許商標庁 (United States Patent and Trademark Office, USPTO)

- 特許出願の全文データを無償で2001年3月15日から提供。対応する特許出願は一年前に日本で公開されている可能性があるので、本研究は2000年以降を対象とする

日本特許庁 (Japan Patent Office, JPO)

- 特許情報の一括ダウンロードサービスを無償で提供。ハードディスクを特許庁へ送る
  - 公開特許公報: 日本語で書かれた通常の国内特許 (パリルートの探索対象)
  - 公表特許公報: 日本特許庁以外を受理官庁とするPCT国際特許を、日本へ国内移行する際に日本語へ翻訳して提出
  - 再公開特許: 日本特許庁を受理官庁とする日本語で書かれたPCT国際特許
- 2021年12月23日に再公開特許の公開を廃止したので、本研究は2021年までを対象とする

欧州特許庁 (European Patent Office, EPO)

- パテントファミリーを含む全世界の特許出願の書誌データベース DOCDB を有償で提供
- 本研究では2022年4月に DOCDB を入手

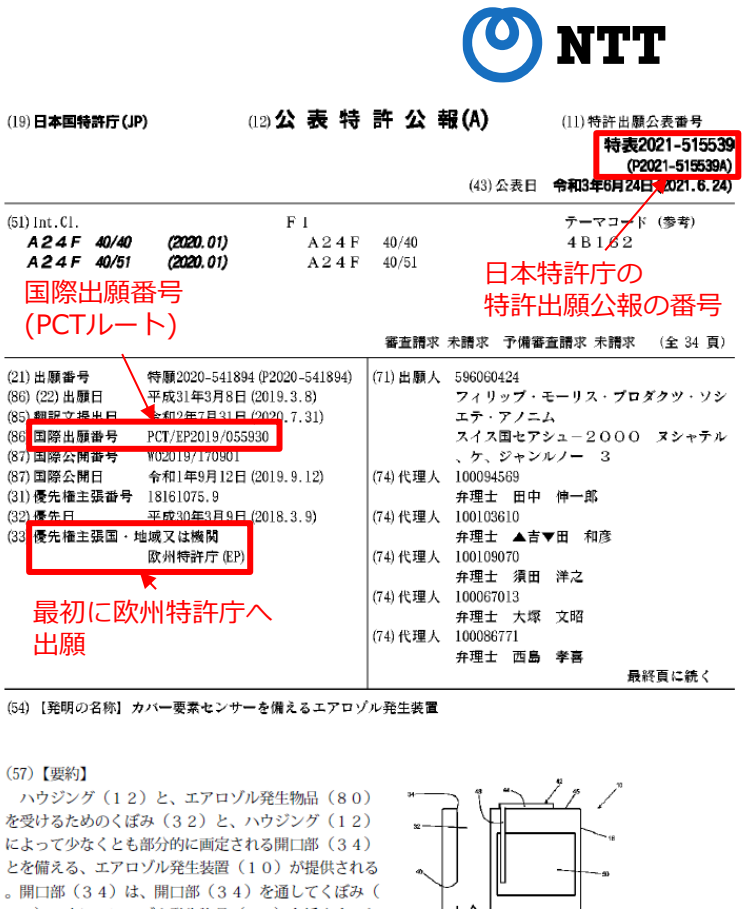
パテントファミリーに基づく日英特許対訳文書対の抽出

網羅性の高い特許対訳文書対の探索ソフトウェアを独自に実装

- DOCDBのパテントファミリー情報を起点として、日本と米国で公開された特許出願の対を求める
- パリルートとPCTルートの両方を探索。日本と米国以外に最初に出願された特許も探索
- 一つの文書をタイトル、要約、本文、請求項の4つの部分に分け、各部分で文対応を計算

国際特許出願制度によって文対応の難しさが違う

- パリルート(優先権主張に基づく出願): 国によりフォーマットが異なり、必ずしも対訳ではない文書対が得られる
- PCTルート: 国際的に統一されたフォーマットが使用され、ほぼ対訳である文書対が得られる



PCTルートの場合、同じ国際出願番号を持つ米国特許庁の特許出願公報を探す

日本特許庁 XMLファイルの例

```
<?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>
```

米国特許商標庁 XMLファイルの例

```
<?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="US20220295684A1-20220922.XML" status="PRODUCTION" id="us-patent-application" country="US" date-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>
```

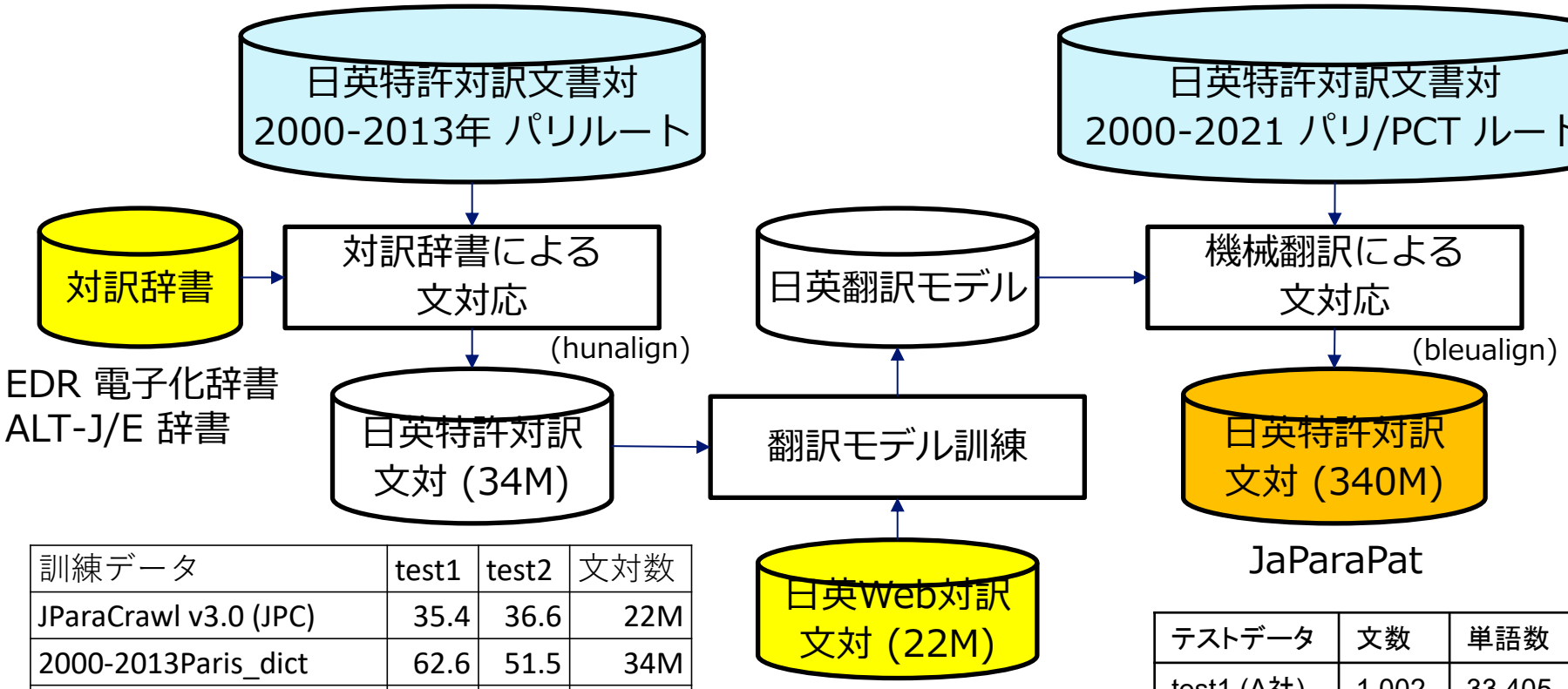
欧州特許庁 DOCDBから変換したXMLデータ

元のXMLはもっと複雑

- Python標準ライブラリの xml.etree.ElementTreeモジュールを使って 必要なデータだけを抽出


```
[
  "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": []
},
```

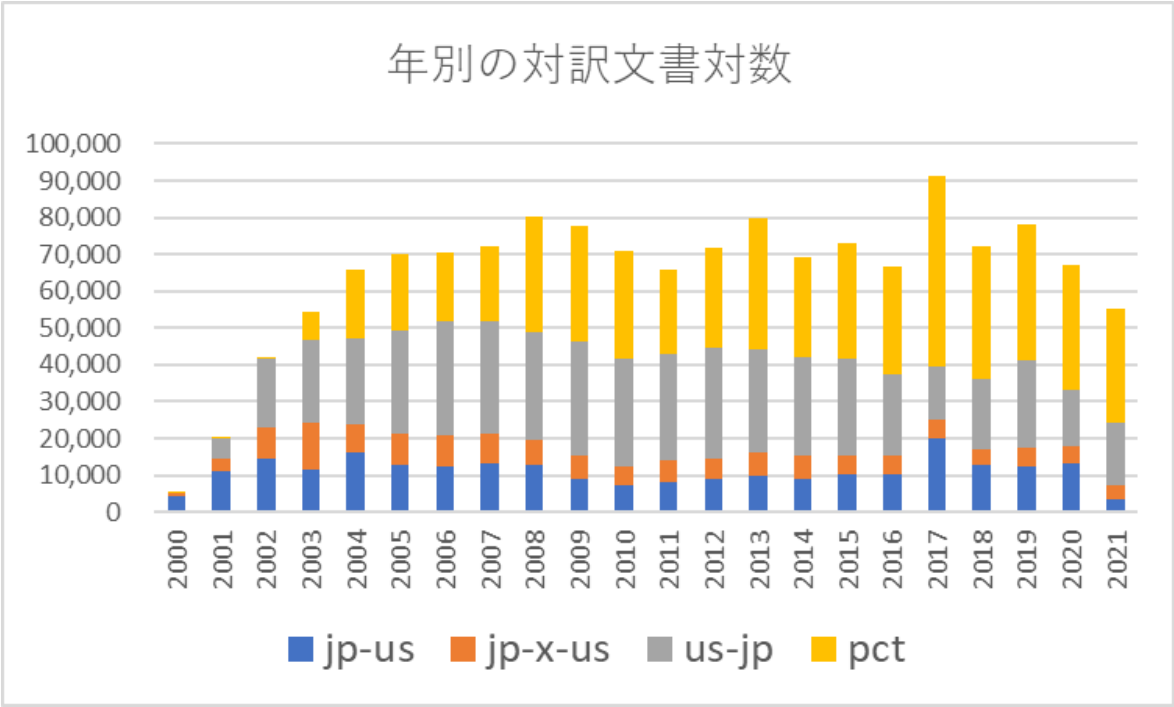
対訳辞書からのブートストラップ



訓練データ	test1	test2	文対数
JParaCrawl v3.0 (JPC)	35.4	36.6	22M
2000-2013Paris_dict	62.6	51.5	34M
2000-2013Paris_dict+JPC	63.6	54.0	56M
2000-2013Paris_trans	63.4	53.0	43M

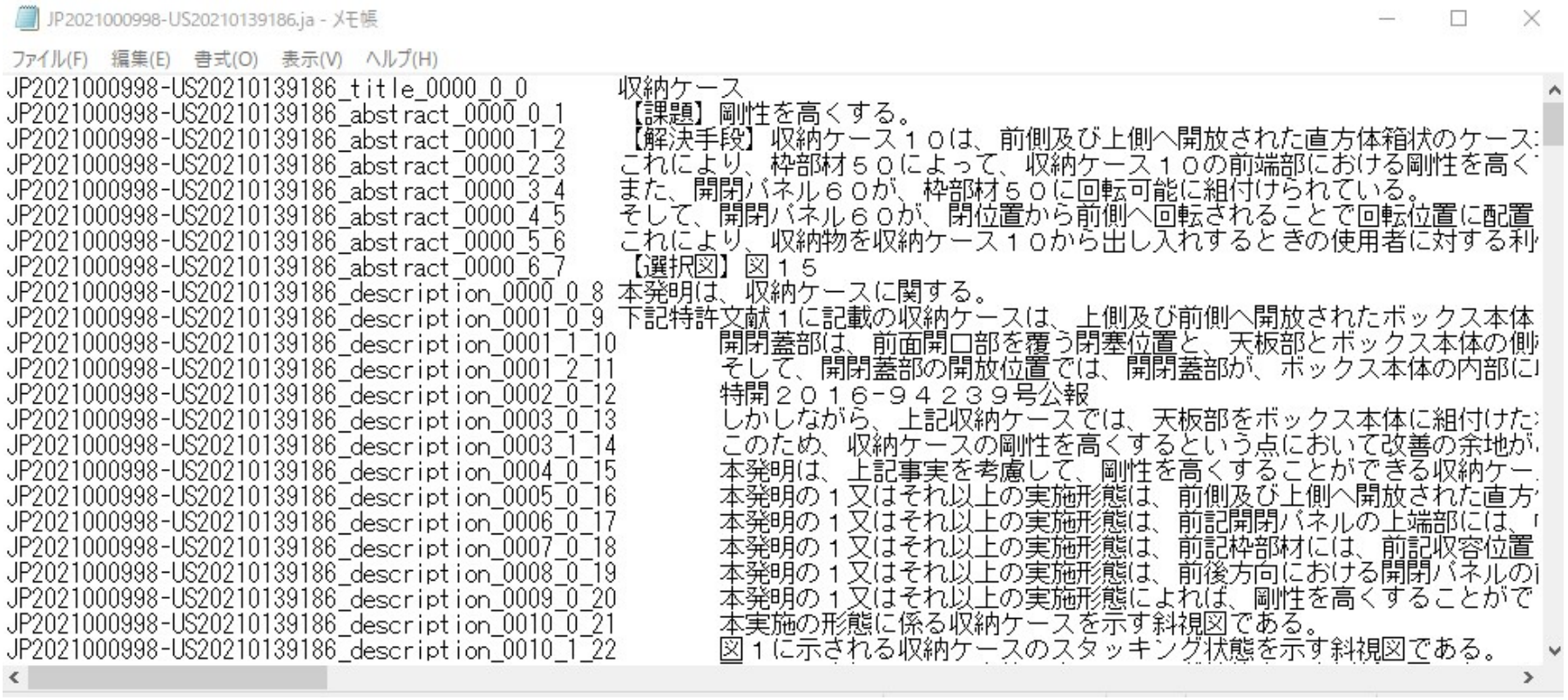
テストデータ	文数	単語数
test1 (A社)	1,002	33,405
test2 (B社)	988	26,945

JaParaPat の概要



ルート	文書数	文数	単語数
パリ	867K	182M	7.4B
PCT	527K	155M	6.2B
合計	1,394K	337M	13.6B

JaParaPat: 日本語文書の例



日本出願番号-米国出願番号_セクション_段落番号_段落内文番号_文番号 ¥t 文

JaParaPat: 英語文書の例

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
```

日本出願番号-米国出願番号_セクション_段落番号_段落内文番号_文番号 ¥t 文

JaParaPat: 文対応と国際特許分類の例

左側: 日本特許と米国特許の文対応

中央: 日本特許のIPC

右側: 米国特許のIPC

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)

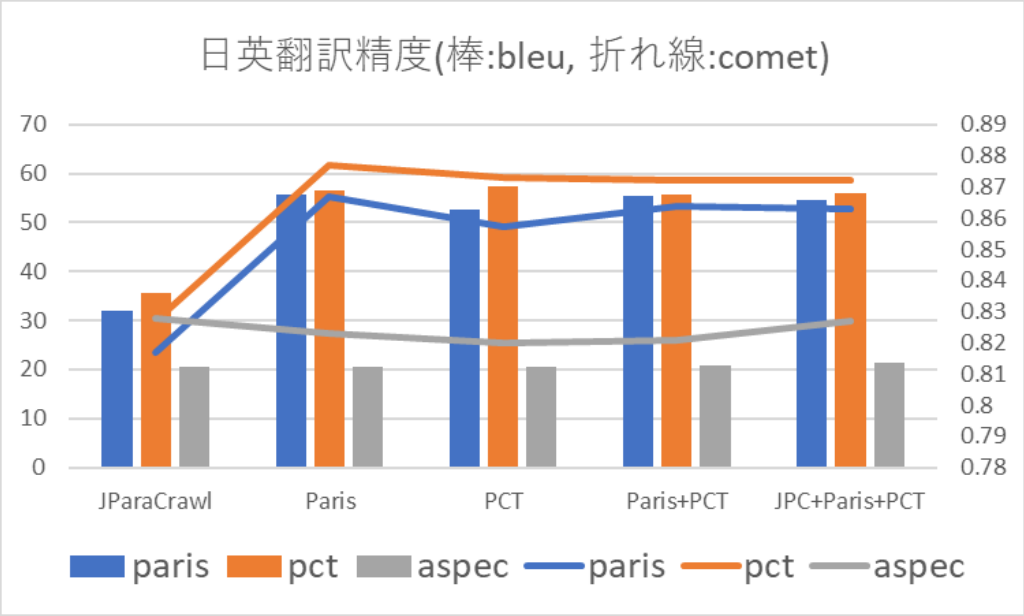
JP2021000998-US20210139186	B65D	5/00
JP2021001262-US20210261830	C09J	11/06
JP2021001957-US20210208529	G03G	15/20
JP2021002617-US20210005717	H01L	29/20
JP2021002715-US20210247578	G02B	6/42
JP2021002959-US20210013564	H01M	10/6563
JP2021002964-US20210009004	B60L	58/12
JP2021002965-US20210008963	B60H	1/00
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

Japanese-to-English Translation Experiment

訓練データ
- パリ/PCTルート
  - 2000年から2021年上半期まで
- JParaCrawl v3.0

テストデータ
- パリ/PCTルート
  - 2021年下半期からランダムに選択
- ASPEC (scientific paper)

翻訳モデル
- Transformer big



Web対訳(JParaCrawl)に比べて
特許対訳(JaParaPat)を使うと
翻訳精度が約20 BLEU ポイント向上

まとめ

日本と米国の特許出願データから得られる日英特許対訳データの量と質を明らかにした

- 2000-2021年の出願データから約100万文書対、3億文対以上の対訳データを収集
- 55 BLEUポイント程度の翻訳精度 (Web対訳に比べ約20ポイント高い)

今後の課題

- 最近の特許を人手で翻訳した特許翻訳のテストセットの作成
  - › 最近の出願データから収集した対訳は、ほとんどが後編集で作成されており、自動評価が混乱する
- 請求項の翻訳
  - › 請求項はとても長く、本文とは文体が異なり、特別な対処が必要
- 日中・日韓対訳データ
  - › 日本特許庁は中国や韓国の特許データも提供