JaParaPat: 大規模日英特許対訳コーパス

永田 昌明　森下 睦　帖佐 克己　安田 宜仁
NTT コミュニケーション科学基礎研究所
{masaaki.nagata,makoto.morishita,katsuki.chousa,norihito.yasuda}@ntt.com

概要

2000 年から 2021 年に日本特許庁 (JPO) と米国特許商標庁 (USPTO) から公開された特許出願から約 3 億文対の日英対訳コーパスを作成した。欧州特許庁 (EPO) が管理する書誌データベース DOCDB からパテントファミリーに基づいて対訳文書対を抽出し、機械翻訳に基づく文対応を用いて対訳文対を抽出した。Web から収集した約 2000 万文対の対訳データ JParaCrawl に対して、特許出願から収集した約 3 億文対の対訳データを追加することにより、日英特許翻訳の精度が約 20BLEU ポイント向上した。

1　はじめに

国際特許出願は非常に数が多いが有限であり、一定の期間を経てすべて公開される。本研究では、日本と米国の国際特許出願から得られる日英特許対訳データの量と質、およびそれらを用いて達成可能な翻訳精度を明らかにすることを目的とする。なお、日本における国際特許出願の翻訳業務の多くは日本語から英語への翻訳であるため、ここでは日英翻訳のみを評価対象とする。

日英特許対訳コーパスを作成する歴史は 20 年近くに及ぶ。内山ら [1] は、日英新聞記事のコンパラブルコーパスから対訳文を抽出する手法 [2] を使って、2007 年に NTCIR-6 特許検索タスクのために約 200 万文対の日英特許対訳コーパスを作成した。これは公開された最初の大規模な日英特許コーパスであり、2008 年に日本語と英語の間の機械翻訳に関する最初の共通タスクである NTCIR-7 特許翻訳タスクで使用された。[3].

特許庁と情報通信研究機構 (NICT) は、日米の特許出願公開公報からパテントファミリーに基づいて約 3.5 億文対の JPO-NICT 英日対訳コーパスを作成した[1)]。このコーパスは、高度言語情報融合フォーラム (ALAGIN) の会員に無償で提供されている。また特許庁は、100 万文対の日英特許対訳 JPO Patent Corpus[2)] を作成し、2015 年に初めて開催されたアジア翻訳ワークショップ (WAT) において特許翻訳の共通タスクに提供した。

JPO-NICT や JPO の特許対訳コーパスは 2015 年頃に対訳辞書に基づく文対応手法 [2] を用いて作成された。一般論として hunalign[4] のような対訳辞書に基づく文対応よりも、Bleualign[5] のような機械翻訳に基づく文対応の方が精度が高いので、最先端の文対応手法を使えば日英特許対訳コーパスの品質を向上できる可能性がある。

近年、海外では大規模な特許対訳コーパスを作成する様々な試みが行われている。COPPA[6] は PCT 出願の抄録から対訳を抽出した。ParaPat[7] は Google Patents の抄録から対訳を抽出した。EuroPat[8] は、USPTO と EPO の特許査定を受けた特許文書から対訳を抽出した。EuroPat は、ParaCrawl プロジェクト [9] で実績がある機械翻訳に基づく文対応を用いている。

我々のアプローチは EuroPat に似ているが、特許査定を受けた特許ではなく未審査の特許出願を使用し、欧州言語間ではなく日英間の対応を付与した。国際特許出願には、パリルートと PCT ルートの 2 つの方法がある。我々の知る限り、本研究は、両ルートの特許出願を広範囲に調査し、タイトル、要約、明細書、特許請求の範囲を含む特許文書のすべての部分に対応を付与する初めての試みである。

2　特許出願データ

2.1　国際特許出願制度

外国で特許を取得する方法には、パリ条約 (Paris Convention) に基づいて直接その国へ出願する方法 (パリルート) と、特許協力条約 (Patent Cooperation Treaty, PCT) に基づく国際出願をその国へ移行する方法 (PCT ルート) がある。

1） https://alaginrc.nict.go.jp/jpo-outline.html

2） http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/

パリ条約ルートでは、ある国で国内出願をした後、パリ条約に基づく優先権を主張して 1 年以内に別の国に出願する。PCT 出願では、PCT 受理官庁 (receiving office) に単一の言語および形式で一つの PCT 出願を行えば、すべての PCT 加盟国で出願日に優先権を確保できる。しかし、ある国で特許権を取得するためには、その国に対して優先日から 30 ヵ月以内に PCT 出願の国内移行 (national phase application) の手続きを行い、その国の国内法に従って特許審査を受ける必要がある。その際にその国の特許庁が受理する言語へ特許出願を翻訳しなければならない。例えば、日本企業が日本語で書いた PCT 出願を世界知的所有権機関 (World Intellectual Property Organization, WIPO) へ提出した場合、日本への国内移行後に日本特許庁 (JPO) は日本語の特許を公開し、米国への国内移行後に米国特許商標庁 (USPTO) は英語の特許を公開する。

2.2 日本特許庁の特許データ

日本特許庁 (JPO) は特許情報の一括ダウンロードサービスを提供している[3)]。企業が利用する場合には会社の登記簿の提出を求められるが、ハードディスクを特許庁に送れば、必要な特許情報を入れて返送してもらえる。

日本の特許公報では、PCT に関する特許出願には通常の国内出願とは別の名前が与えられている。公開特許公報 (published pattent application) は、日本語で書かれた通常の国内特許である。これがパリ条約に基づく優先権主張の探索対象となる。公表特許公報 (Japanese translation of PCT international patent application) は、JPO 以外を受理官庁とする国際特許出願を、日本へ国内移行する際に提出された日本語への翻訳である。

再公開特許 (domestic re-publication of PCT international patent application) は JPO を受理官庁とする日本語で書かれた国際特許出願である。JPO は、再公開特許を公開する制度を、2021 年 12 月 23 日に廃止した。この日以降、最初に日本へ日本語で出願された PCT 出願は、一定の補正を経て特許と査定された場合にのみ利用可能となる。そのため本研究の対象期間は 2021 年までとした。

我々は、対訳データを作成するために、発明のタイトル、要約、明細書、請求項に対応する XML 要素の p タグで囲まれたテキストを抽出した。すなわち、請求項番号、段落番号、数式、図などは抽出していない。日本の特許出願は 2004 年 1 月から XML 形式になっているが、2004 年以前は SGML 形式であった。SGML 形式のデータは、基本的に XML 形式と抽出対象が同じになるようにした。

2.3 米国特許商標庁の特許データ

米国特許商標庁 (USPTO) は、特許出願の全文データを提供している。[4)] 説明書と文書型定義 DTD は USPTO の Web ページにある。[5)] USPTO は、2001 年 3 月 15 日から特許出願の全文データを提供している。対応する特許出願は、米国で公開される 1 年前に日本で公開されている可能性があるため、本研究の対象期間は 2000 年からとしている。

USPTO の PCT 特許出願には、JPO の公表特許公報と再公開特許のような区別がないので、PCT/JP2005/003817 のように文書番号が PCT で始まる場合を PCT 出願とみなした。

2.4 欧州特許庁の書誌データ

欧州特許庁 (EPO) は、全世界の特許書誌データ (DOCDB) を有償で提供している。DOCDB のサンプル[6)] とマニュアル[7)] は EPO の Web サイトからダウンロードできる。

我々は、パテントファミリーの情報を取得するために DOCDB を入手した。パテントファミリーとは、一つの発明を保護するために様々な国で取得された特許の集合である。パテントファミリーは、DOCDB の優先権主張のデータを解析することにより得られる。DOCDB の XML では、優先権主張の対象となっている文書の kind-code から通常の特許 (A) と PCT 出願 (W) を区別できる。

3 手法

3.1 文書対応

EPO の DOCDB から得られるパテントファミリーに基づいて、JPO と USPTO から公開された特許

3) https://www.jpo.go.jp/system/laws/sesaku/data/download.html

4) https://developer.uspto.gov/product/patent-application-full-text-dataxml

5) https://www.uspto.gov/learning-and-resources/xml-resources

6) https://www.epo.org/searching-for-patents/data/bulk-data-sets/docdb.html

7) https://www.epo.org/searching-for-patents/data/bulk-data-sets/manuals.html

出願を対応づける。元データはすべて XML であり、Python 標準ライブラリの xml.etree.ElementTree モジュールを使用して以下に記述する文書対応の手順を実装した。同じパテントファミリーに所属する日本語と英語の特許出願の対の中で一番古いものを互いに対訳である文書対とみなす。

特許対訳文書対の探し方は、パリルートと PCT ルートで若干異なっている。パリルートの場合、基本は、一方が他方を優先権主張の対象とする場合である。本研究では、日本に出願した特許を優先権主張の対象とする米国の特許との対を 'jp-us'、米国に出願した特許を優先権主張とする日本の特許との対を 'us-jp' と呼ぶ。さらに例えば最初に中国で出願した特許を日本と米国へ出願する場合のように、共通の特許を優先権主張の対象とする日本の特許と米国の特許の対も抽出対象とし 'jp-x-us' と呼ぶ。

PCT ルートの場合、日本 (JPO) は再公表特許と公表特許公報の出願番号を求め、米国 (USPTO) は文書番号が PCT で始まるものを求める。両方が DOCDB に存在し、かつ、一致した場合、日本の特許出願と米国の特許出願を互いに翻訳な文書対とみなす。

3.2 文対応

日本の特許出願と米国の特許出願は、タイトル、要約、明細書、請求項に分割し、それぞれのセクションごとに文対応を行った。文分割は、日本語、英語ともに moses の split-sentences.perl [8] を用いた。

まず対訳辞書に基づく文対応 [2] を用いて最初の特許対訳データを作成した。次にこの特許対訳データと Web から収集した大規模日英対訳対データ JParaCrawl[10] を用いて日本語から英語への翻訳モデルを作成した。そして機械翻訳に基づく文対応 [5] を用いてもう一度特許対訳データを作成した。

対訳辞書に基づく文対応は、日本語文と英語文の対において、対訳辞書に収録されている対訳語句がそれぞれに出現する割合をこの文対の類似度の尺度とする。対訳辞書としては、EDR 対訳辞書[9] を使用した。日英対訳語句対の数は 1,690,174 件である。日本語の単語分割には mecab-unidic [10]、英語のトークナイズには TreeTagger [11] を用いた。機械翻訳に基づく文対応には Bleualign[12]、機械翻訳には fairseq [11] を用いた。日本語文書を英語へ翻訳し、BLEU を文の類似度の尺度として動的計画法により文対応を求めた。

4 実験

4.1 特許対訳データ

表 5 に日本の特許出願の公開年に基づいて 2000 年から 2021 年まで年別に収集された対訳文書対と対訳文対の数を示す。合計で約 140 万の対訳文書対から約 3.5 億の対訳文対を収集した。jp-us, jp-x-us, us-jp, pct は 3.1 節に述べた出願ルートと最初の出願国に基づく分類である。

表 1 訓練データの文書数、文数、英語側の単語数

ルート	文書数	文数	単語数
Paris	866,931	181,907,843	7.4B
PCT	527,068	154,860,596	6.2B
Paris+PCT	1,393,999	336,768,439	13.6B

表 2 テストデータの文数と英語側の単語数

テストデータ	文数	単語数
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

表 3 文対応手法の比較

訓練データ	test1	test2	文対数	更新
JParaCrawl	36.4	36.6	22M	19K
2000-2013Paris_dict	62.6	51.5	34M	17K
2000-2013Paris_dict+JPC	63.6	54.0	56M	26K
2000-2013Paris_trans	63.4	53.0	43M	16K

4.2 実験条件

日英特許対訳コーパス JaParaPat の品質を確認するため、日本語から英語への翻訳実験を行った。2000 年から 2021 年の前半までの対訳データを訓練データとした。表 1 に訓練データの文書数、文数、英語側の単語数を示す。

表 2 にテストデータの文数と英語側の単語数を示す。開発データとテストデータは、2021 年の後半 (second half) のパリルートと PCT ルートからそれぞ

8) https://github.com/moses-smt/mosesdecoder/blob/master/scripts/ems/support/split-sentences.perl
9) https://www2.nict.go.jp/ipp/EDR/ENG/indexTop.html
10) https://taku910.github.io/mecab/
11) https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
12) https://github.com/rsennrich/Bleualign

表 4　様々な訓練データに対する日英翻訳の精度

訓練データ	Paris		PCT		test1		test2		ASPEC		文対数	更新数
	bleu	comet	bleu	comet	bleu	comet	bleu	comet	bleu	comet		
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

れ 2000 文と 1000 文をランダムに選択した。さらに 2022 年以降に公開予定の社内の日本語 PCT 出願を特許翻訳を専門とする二つの翻訳会社が英語へ翻訳したものをテストデータとして用意した。対象領域は情報通信 (ICT) に関するもので、ハードウェアからソフトウェアまで幅広い内容を含んでいる。また特許対訳データの科学技術論文への適用可能性を調べるために、ASPEC[12] のテスト文も使用した。WAT-2023 の特許翻訳に関する共通タスクで使用されている JPO Patent Corpus のテストセットは最も新しいものでも公開年が 2019-2020 であり、収集期間に含まれているので使用しなかった。

機械翻訳ソフトウェアは fairseq [11] を使用し、翻訳モデルは Transformer big [13] を使用した。Transformer のハイパーパラメタを表 6 に示す。訓練データおよびテストデータのトークナイズは、sentencepiece [14] を使用した。特許対訳データから 7M 文対、JParaCrawl から 3M 文対をランダムにサンプリングして、sentencepiece のモデルを訓練した。語彙数は、日本語と英語ともに 32K である。翻訳の精度は sacreBLEU [15, 16] と COMET(wmt22-comet-da) [17] で評価した。特許翻訳では専門用語を正しく訳出できることが重要なので、本研究では BLEU を主たる評価尺度とする。

4.3　文対応法の比較

最初に二つの文対応手法について評価した。Web から対訳データを収集する研究では、対訳辞書を用いた文対応よりも翻訳器を用いる文対応の方が高品質な対訳文が得られると報告されている。[9, 10]

まず 2000 年から 2013 年のパリルートの日英対訳文書対から、対訳辞書に基づく文対応で約 34M 文対の対訳データ (2000-2013Paris_dict) を作成した。次にこの対訳文対と JParaCrawl(2000-2013Paris_dict+JPC) から翻訳モデルを作成し、機械翻訳に基づく方法で約 43M 文対の対訳データ (2000-2013Paris_trans) を作成した。それぞれの翻訳モデルについて、特許テスト文に対する翻訳精度を表 3 に示す。[13)]

テストセットによる翻訳精度の違いはさておき、表 3 からは、Web から収集した対訳データよりも、特許出願から収集した対訳データを使用する方が翻訳精度が大幅に高いこと、特許出願から収集した対訳データと Web から収集した対訳データを合わせると、特許翻訳の精度が少し向上すること、対訳辞書に基づく文対応よりも機械翻訳に基づく文対応の方が多くの対訳データを収集できること (34M から 43M)、そして翻訳精度から判断すると機械翻訳に基づく文対応により得られる対訳データの方が高品質であることが分かる。

4.4　日英翻訳精度

2000 年から 2021 年のパリルート (182M 文対) と PCT ルート (155M 文対) および JParaCrawl(22M 文対) から訓練した翻訳モデルの翻訳精度を表 4 に示す。Web 対訳データである JParaCrawl と比較すると、特許対訳データを使用することにより、特許テスト文の翻訳精度は大幅に向上する。パリルートと PCT ルートを比較すると、パリルートのテスト文はパリルートの対訳データの方が翻訳精度が高く、PCT ルートのテスト文は PCT ルートの対訳データの方が精度が高い。JParaCrawl とパリルートと PCT ルートを合わせる (JPC+Paris+PCT) と、JParaCrawl だけ (JPC) に比べて日本語から英語への特許翻訳の精度が約 20BLEU ポイント向上した。

5　おわりに

2000-2021 年の日本と米国の特許出願から網羅的に特許対訳データを収集し、約 350M 文対の特許対訳コーパス JaParaPat を作成した。訓練データの量が Transformer big モデルで学習できる限界に達している可能性があるので、今後は機械翻訳のスケール則に関する研究 [18, 19] を参考にして、モデルのパラメタを増やす検討が必要である。

13)　この時点では 2021 年の対訳データは存在しないので、社内のテストデータを使用した。

参考文献

[1] Masao Utiyama and Hitoshi Isahara. A Japanese-English patent parallel corpus. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark, September 10-14 2007.

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

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

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

[5] Rico Sennrich and Martin Volk. 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, October 31-November 4 2010. Association for Machine Translation in the Americas.

[6] Marcin Junczys-Dowmunt, Bruno Pouliquen, and Christophe MazencAndrew. 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*, pp. 15–19, 2016.

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

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

[9] 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. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics.

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

[11] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 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)*, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

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

[14] Taku Kudo and John Richardson. 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*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

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

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

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

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

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

A 年別の対訳文書対と対訳文対の数

表 5 抽出された対訳文書対と対訳文対の数 (2000 年から 2021 年)

	対訳文対				対訳文書対			
	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
合計	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			

B 翻訳モデルのハイパーパラメタ

表 6 Transformer のハイパーパラメタ

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