平成２８年度
ベトナム・タイ語の対訳コーパス・辞書の
自動作成に向けたツール等の検証調査

調査報告書

平成２９年３月

一般財団法人　日本特許情報機構

目次

１．概要

１．１ 目的

企業がグローバルに経済活動を展開するにあたり、海外の特許情報の重要性が急速に高まってきている。特に東南アジア諸国連合(以下、ASEAN)については、中国一国集中リスクの回避や経済連携協定(EPA)締結を背景に事業進出への関心が高まっており、ASEAN 諸国の特許情報へのアクセス性の向上が求められている。一方で、ASEAN 諸国の言語は日本語との類似性が低く、また十分な量のテキストデータの入手が困難であることも一因となり、ASEAN 言語と日本語を相互に翻訳するための機械翻訳の研究や実用化は英日や中日、韓日翻訳と比べ遅れている。さらには、ASEAN 諸国の特許公報は本文のテキストデータが利用不可能であるケースも見られ、特許情報の把握には多大なコストを要するのが現状である。

こうした状況を踏まえ、本調査では、機械翻訳の活用による ASEAN 諸国の特許情報へのアクセス性向上に向けて、ASEAN 言語の機械翻訳に必要な高精度な対訳コーパス及び辞書の作成方法に関する調査を行う。具体的には、調査対象言語としてベトナム語とタイ語の二つを選定し、将来的にこれらの言語の高精度な対訳コーパス及び辞書を機械的に大規模に構築する可能性を検証するために、両言語にかかる言語資源の作成に必要となるツールの調査と作成方法を検討することを目的としている。

１．２ 調査概要

前節で述べた目的を達成するために、次の 2 項目について調査を行う。

・対訳コーパス・辞書作成ツールの調査

・パテントファミリー以外から対訳コーパス・辞書を作成する方法の検討

以下では、それぞれの項目に関し、調査内容と実施方法の概要を述べる。

（1）対訳コーパス・辞書作成ツールの調査

既に特許庁にて大規模な英日、中日、韓日対訳コーパスや対訳辞書を開発しているように、特許向けの対訳コーパスや対訳辞書を大規模かつ効率的に開発するためには、異なる言語で同内容の出願がなされているパテントファミリーから、翻訳関係にある文を抽出して対訳コーパスを作り、さらにその対訳コーパスを用いて専門用語と相手方言語の訳語を抽出し対訳辞書の候補とする処理が必要である。

そこで本調査においては、ベトナム語及びタイ語の対訳コーパスと辞書の作成においてもそれらの場合と同様パテントファミリーを利用した方法が実施可能であるかを検証する。その方法を採用するために、OCR ツール[1]、文分割ツール、トークナイザ、文アライメントツール、フレーズテーブル作成ツール等としてどのようなツールが存在するかを調査し、

[1] ベトナム語の公報データは、登録特許は DigiPat で OCR 処理されたテキストが入手可能である。そのため、OCR ツールで公報イメージを読み取ってテキスト化した上で評価を行うのは本調査ではタイ語だけとする。

以下の観点で整理する。

・ツールの利用条件
・ツールの入手方法
・ツールの利用環境の構築方法、使用方法
・ツールのコスト及びツールと依存関係にある他のソフトウェアのコスト

また上の調査結果をもとに選定したツールを用いてベトナム語とタイ語の明細書等のデータを処理して、その有用性を明らかにする。

下の図は、泰日対訳コーパス・辞書を作成する場合を例に、パテントファミリーから対訳コーパスならびに対訳辞書を作成する過程の概略と、その過程の各処理で用いるツールを示したものである[2]。

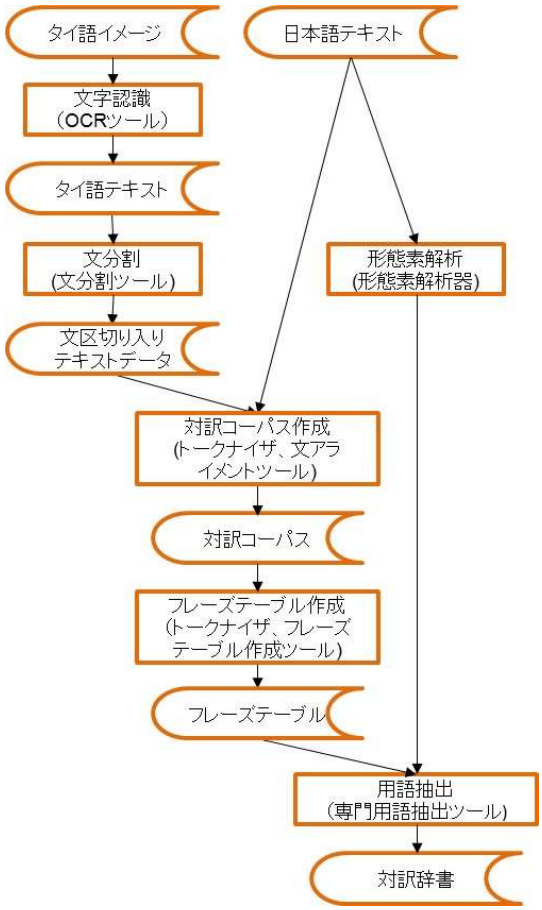


図 1.2-1 パテントファミリーから対訳コーパス・対訳辞書を作成する処理の概略
（処理を示す矩形内のテキストにおいて、カッコ内が使用するツールを示す）

[2] 対訳コーパス作成のフェーズにおいて、使用するツールとしてトークナイザも表示しているが、文アライメントツールの使い方によっては、トークナイザを使わずに行うことも可能である。

（２）パテントファミリー以外から対訳コーパス・辞書を作成する方法の検討

特許向け機械翻訳の訓練で用いる対訳コーパスや対訳辞書は（１）で述べたようにパテントファミリーを用いて作成することが望ましいが、現時点で利用可能なデータだけでは十分な量が確保できない場合には、それ以外のデータも利用しなければならない可能性がある。そこでパテントファミリーのデータを補う目的で、パテントファミリー以外の言語資源を用いて対訳コーパス及び辞書を試作する方法について調査する。

具体的には以下の作成方法について調査する。

①英語を中間言語とした辞書作成方法

②英語を中間言語とした対訳コーパス作成方法

③TED からの対訳コーパス・辞書作成方法

④Wikipedia からの対訳コーパス・辞書作成方法

⑤人手翻訳による対訳コーパス・辞書作成方法

⑥他事業の報告書に記載された言語資源を利用した対訳コーパス・辞書作成方法

⑦その他の言語資源を利用した対訳コーパス・辞書作成方法

調査では、法的・技術的に作成可能かを調べ、可能性のある方法については、さらに以下の観点で調査を行う。

・作成可能な対訳コーパス・辞書の分量
・作成される対訳コーパス・辞書の特徴
・作成の具体的手順（作成の準備含む）
・作成に要する期間（作成の準備含む）
・作成コスト
・作成に際して存在する課題

１．３ 調査結果概要

１．３．１ 対訳コーパス・辞書作成ツールの調査

（１）OCR ツールの調査

ベトナム語については、既に OCR ツールを使って読み取り済みのテキストを公報のイメージデータと比較して性能を評価した。明細書 30 件(計 655 ページ)分のデータで評価したところ、正解率は 99.66%であった。このテキストは DigiPat で公開されているもので、OCR での読み取り後に人手によるチェックを実施しているかは不明であるが、この精度であれば、これ以上の人手チェックは行わずとも対訳コーパスに利用できる精度であると言える。

タイ語の OCR ツールについては２つのツールについて調査した。明細書 20 件(計 228 ページ)の読み取りを実施し性能を評価したところ、正解率は一つが 96.65%、もう一つが 92.37%であった。両者とも 90%を超え実用レベルであるとは言えるが、読み取り結果を人手

によりチェックすることで、読み取り結果の確認・修正を行うことが好ましい。

（２）文分割ツールの調査

ベトナム語用の文分割ツールとしては、英語の特許文書における文分割用に開発された機械学習ベースの文分割ツールと、前処理機能として文分割機能を有するベトナム語トークナイザの二つの性能を 7,473 文のテキストを用いて調査した。両者の正解率はそれぞれ 94.9%と 92.8%であり、英語向けの文分割ツールの方が高い性能を示した。

タイ語用の文分割ツールとしては、OpenNLP というツールを評価した。OCR ツールの評価の過程で作成したタイ語のテキスト 3,018 文を用いて文分割ツールの性能を確認したところ、正解率は 15.4%となり、他の言語における文分割ツールの精度と比べ著しく低いことが分かった。これは、英語のピリオドや日本語の句点のような文末記号がタイ語には存在しないことが最も大きな原因であると考えられる。

（３）トークナイザの調査

トークナイザについては、コーパス作成ツールや辞書作成ツールを使用する中で用いるツールであるため、トークナイザ単独での大規模な評価は実施しなかったが、ベトナム語のトークナイザについては以下の知見が得られた。評価したのは vnTokenizer と JVNSegmenter の二つである。処理精度は vnTokenizer の方が良いと判断できたが、同ツールは、評価に使用した 30 万文のテキストの処理で異常終了したことと、特定のテキストについては非常に処理時間がかかるため、大規模なテキスト処理に使用するには注意が必要である。

（４）コーパス作成ツールの調査

コーパス作成ツールとして内山らのツールを評価した。このツールは、これまでの特許庁の各種事業において英日や中日対訳コーパス・辞書の開発で使用されてきたツールであり、ベトナム語、タイ語と日本語、英語間の文アライメントにも使用できることが確認できた。ただし、各言語対での文アライメントの評価でスコアが A（文対の文の内容がほぼ一致している)と判断された割合は以下のとおりであり、特許庁の過去の調査事業における中日機械翻訳用辞書開発に関する調査[3]で大規模に実施された中日文アライメントでの精度約 90%と比べると、越英を除き、やや低めの精度となっている。

越日： 72.0%
越英： 92.0%
泰日： 86.0%
泰英： 79.5%

[3] 平成 25 年度「中国特許文献の機械翻訳のための新語に関する調査」

（５）辞書作成ツールの調査

辞書作成ツールとして、フレーズテーブルの作成にはMosesとmgizaを用い、フレーズテーブルから抽出する見出しの選定には言選を用いた。これらのツールも、これまでの特許庁の各種事業において英日や中日対訳コーパス・辞書の開発で使用されてきたツールであり、英越辞書登録候補語の抽出精度が他の言語対よりもやや低いものの、実用レベルの精度であると判断できる結果が得られた[4]。各言語対での辞書登録候補語の評価で、最大3個の訳語を抽出した中で第一訳語のスコアがA（両言語のフレーズ・語の意味がおおむね一致している）と判断された割合は以下のとおりであった。

日越： 86.5%
英越： 67.5%
日泰： 82.5%
英泰： 74.0%

１．３．２ パテントファミリー以外から対訳コーパス・辞書を作成する方法の検討

個々の作成方法に関する詳細は3章で述べるが、インターネット上で公開されている言語資源を利用して作成する対訳コーパスは、利用条件として特別な条件が明示されていない限り、コーパスそのものを公開したり第三者に販売したりせず、あくまで統計翻訳の学習目的で利用するのであれば著作権上の問題はないと考えられる。また、作成された対訳コーパスを利用して辞書登録候補語を抽出した後に人手でチェックすることで作成した対訳辞書についても法的な問題はないと考えられるが、既存の対訳辞書を利用する場合やWikipedia の言語間リンクの見出しから直接辞書を作成するような場合には当然のことながらライセンス契約や使用に際してのクレジット表示など適切な法的措置を講じなければならない場合もあるので注意が必要である。

法的な課題がクリアされたとして、それぞれの方法で対訳コーパスや辞書を低コストで作成する上での最大の問題は、特許翻訳用の機械翻訳システムの学習や対訳辞書の開発に利用するのに適していると考えられる特許分野もしくは科学技術分野の大規模な言語資源が少ないことである。例えば、Wikipediaの言語間リンクの見出しから直接辞書を作成する手法では、対訳辞書として越日辞書16.3万語の辞書が作成できるが、これらには人名や地名、会社名、エンターテインメントの作品名など技術用語以外の見出しが多く含まれ、特許分野のテキストに対するトークナイズや文アライメントの処理に用いるには適しているとは言い難い。今回の調査で有用性が高いと判断できたものは次の二つである。

・OPUS(Open Parallel Corpus)のサイトで公開されているソフトウェアのテキスト
・IPC分類表

また、人手翻訳による対訳コーパス・辞書作成は、高い品質の対訳コーパスが構築できる

[4] 平成25年度「中国特許文献の機械翻訳のための新語に関する調査」では100万語規模で実施された中日対訳辞書候補抽出の精度が約70%であったと報告されている。

が翻訳コストが非常に高くなるため、すぐに実行するのは現実的には難しいと考えられる。

２．対訳コーパス・辞書作成ツールの調査

本章では、ベトナム語及びタイ語の対訳コーパスや辞書、すなわち、それらの言語と日本語・英語間のコーパスである越日／越英／泰日／泰英対訳コーパス、ならびに、対訳辞書を作成するためのツール（OCR ツール、文分割ツール、トークナイザ、アライメントツール等）を調査するとともに、利用可能なツールについて試用した結果について述べる。

２．１　ベトナム語用ツールの調査

２．１．１　OCR ツールの調査

（１）利用可能なツールの調査

ベトナム語の権利化された特許は、OCR でテキスト化された明細書全文（以下 OCR テキスト）が DigiPat[5]に掲載されている。そのため、ベトナム語用の OCR ツールとしてどのようなツールが存在するかの調査と OCR ツールを用いた読み取り作業はいずれも実施しなかった。

（２）評価

（２－１）評価で使用したデータ

DigiPat で OCR テキストと明細書イメージデータが入手可能な案件の中で日本語と英語でも出願されたパテントファミリーを持つ案件計 30 件を候補として選定した。左記 30 件のうち､15 件は日本の法人が出願したもの､残り 15 件は北米の法人が出願したものである。また、読み取り精度の評価にあたって、特定の公報だけ極端に文字数が多くなるのを避けるため、それぞれの案件において評価対象とするページ数は最大 27 ページとし、30 件全体でのページ数は 655 ページとした。平均では一公報あたり 21.8 ページである。なお、30 件全体でのワード数(音節数)は 251,599 である。

（２－２）評価手順

OCR ツールを用いた読み取り作業は行わないので、評価の手順は以下のようになる。

①OCR テキストにおいて、本来の読み取り対象であるテキスト以外に一緒に読み取られてしまったページ番号や登録番号等のノイズを削除する。

②ベトナム語の専門家が OCR テキストを明細書イメージデータと見比べて、OCR テキストにおいて誤認識された文字を特定し、正解率を算出する。

（２－３）評価結果

表 2.1-1 に文献ごとの OCR テキストの正解率を示す。全データに対しての正解率は 99.66％であった。

[5] http://digipat.noip.gov.vn/

表 2.1-1 OCR 正解率（ベトナム語）

#	出願番号(VN)	IPC	作業頁数	文字数(正解)	誤認識文字数	OCR正解率
1	1-2007-01812	G11B	27	41,345	258	99.38%
2	1-2005-01787	F02B	18	25,939	9	99.97%
3	1-2005-00402	A61F	27	44,995	246	99.45%
4	1-2008-02431	C07C	27	12,472	5	99.96%
5	1-2008-02720	C07D	26	31,715	28	99.91%
6	1-2008-00617	B24B	23	31,266	148	99.53%
7	1-2008-00182	B65D	15	21,305	20	99.91%
8	1-2004-00924	A61K	21	27,736	444	98.40%
9	1-2008-01287	G03G	27	36,966	1,118	96.98%
10	1-2005-00564	E02D	11	13,387	17	99.87%
11	1-2012-03600	H02M	21	25,680	7	99.97%
12	1-2012-03893	G11B	27	32,888	50	99.85%
13	1-2012-03659	H04W	18	27,393	2	99.99%
14	1-2013-00360	F04C	9	10,351	14	99.86%
15	1-2012-02596	C08J	27	39,492	58	99.85%
16	1-2005-01458	H04L	20	32,752	9	99.97%
17	1-2006-00845	H01J	21	29,644	14	99.95%
18	1-2002-00889	C01B	27	37,151	50	99.87%
19	1-2005-01263	A61K	26	42,048	36	99.91%
20	1-2006-00339	C07D	15	21,192	29	99.86%
21	1-2007-00963	A61K	24	32,757	55	99.83%
22	1-2004-01292	G03F	23	17,945	6	99.97%
23	1-2004-01208	B32B	27	44,447	135	99.70%
24	1-2005-00810	B01D	19	28,523	61	99.79%
25	1-2006-02057	G02B	26	46,309	188	99.59%
26	1-2012-02960	C08J	27	42,754	13	99.97%
27	1-2012-01489	E02F	27	39,840	55	99.86%
28	1-2012-00180	D01D	19	25,011	29	99.88%
29	1-2011-03350	F27D	14	23,284	26	99.89%
30	1-2011-02026	G08B	16	23,885	8	99.97%
計			655	910,472	3,138	99.66%

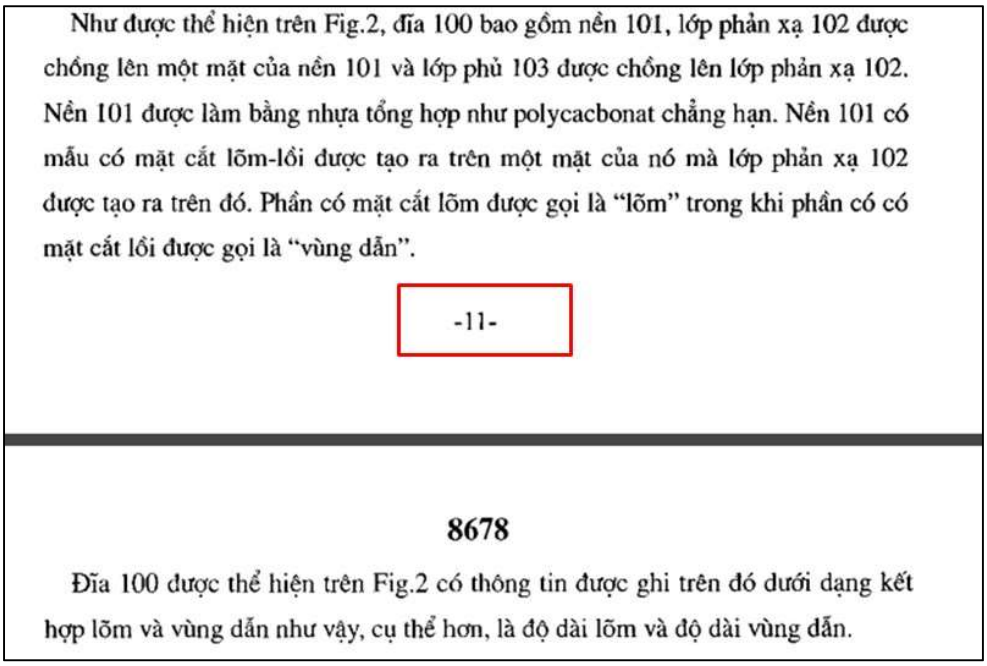
（3）考察

DigiPat の OCR テキストが OCR ツールによる認識結果に対してどの程度人手による修正がなされているのかは不明であるが、当該テキストを対訳コーパスや対訳辞書を作成する目的で使用するにあたっては以下の点で注意する必要があると考えられる。

（３－１）ノイズの除去

文献により、ページ上部にプリントされた登録番号やページ下部のページ番号が適切に除去されている文献もあれば、除去されていない文献もあった。これらの番号は OCR テキストを検索目的で使用する場合にはほとんど問題にはならないが、機械翻訳のための言語資源として利用することを考えた場合、このような不要な語句が残っていると、それをもとに作成した対訳コーパスで学習した統計翻訳(SMT)エンジンで翻訳を行うと、原文に出現しない数字が訳出されてしまう、いわゆる湧き出しの原因となる。そのため（２－２）の手順①で示したように、これらの登録番号やページ番号は削除する必要がある。今回の調査では、これらが残っていた文献の件数は 22 件で、全文献数の 73%であった。

以下に、ノイズが含まれていた公報イメージと、このイメージから読み取られたノイズを含んだ読み取り結果の一例を示す。


Như được thể hiện trên Fig.2, đĩa 100 bao gồm nền 101, lớp phản xạ 102 được chồng lên một mặt của nền 101 và lớp phủ 103 được chồng lên lớp phản xạ 102. Nền 101 được làm bằng nhựa tổng hợp như polycacbonat chẳng hạn. Nền 101 có mẫu có mặt cắt lõm-lồi được tạo ra trên một mặt của nó mà lớp phản xạ 102 được tạo ra trên đó. Phần có mặt cắt lõm được gọi là "lõm" trong khi phần có có mặt cắt lồi được gọi là "vùng dẫn".

-11-

8678

Đĩa 100 được thể hiện trên Fig.2 có thông tin được ghi trên đó dưới dạng kết hợp lõm và vùng dẫn như vậy, cụ thể hơn, là độ dài lõm và độ dài vùng dẫn.

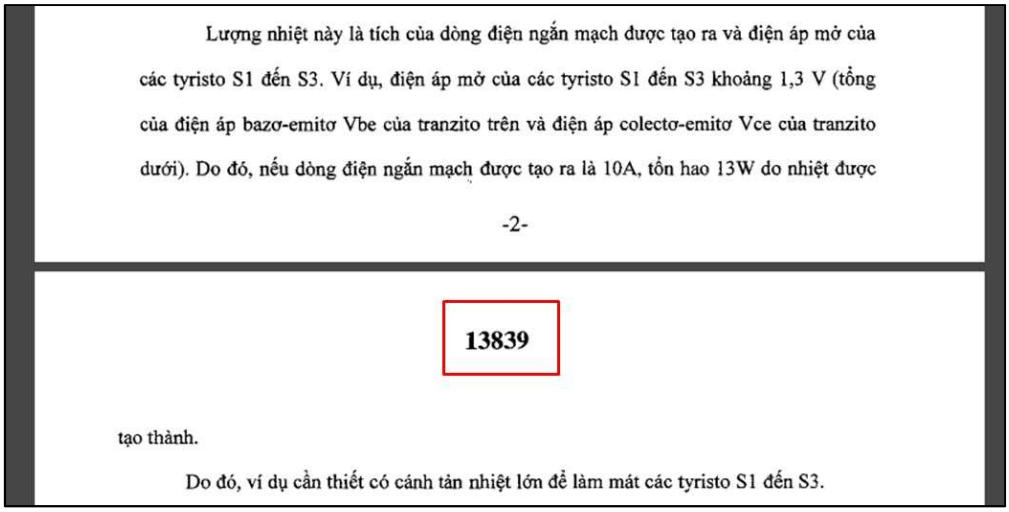
図 2.1-1 ベトナム語公報のイメージデータの一例（その 1）
（赤枠の中がノイズとなって残っているページ番号）

上のイメージに対応する OCR テキストは以下のように赤枠のページ番号部分が「-li-」（英小文字エル、英小文字アイ）となっていた。同じ公報のこのページ以外ではページ番号は削除されていたので、「-数字-」というパターンで読み取り後に機械的に削除する処理を行ったものの、li は数字ではないため、この処理による削除対象から漏れてしまったと考えられる。

Như được thể hiện trên Fig.2, đĩa 100 bao gồm nền loi, lớp phản xạ 102

được chồng lên một mặt của nền loi và lớp phủ 103 được chồng lên lớp phản xạ 102. Nền loi được làm bằng nhựa tổng hợp như polycacbonat chẳng hạn. Nền loi có mẫu có mặt cắt lõm-lồi được tạo ra trên một mặt của nó mà lớp phản xạ 102 được tạo ra trên đó. Phần có mặt cắt lõm được gọi là ″lõm″ trong khi phần có có mặt cắt lồi được gọi là ″vùng dẫn″. -li-
Đĩa 100 được thể hiện trên Fig.2 có thông tin được ghi trên đó dưới dạng kết hợp lõm và vùng dẫn như vậy, cụ thể hơn, là độ dài lõm và độ dài vùng dẫn.

また下の図は、ページ上部の登録番号が残っていた事例における公報イメージの一例である。


Lượng nhiệt này là tích của dòng điện ngắn mạch được tạo ra và điện áp mở của các tyristo S1 đến S3. Ví dụ, điện áp mở của các tyristo S1 đến S3 khoảng 1,3 V (tổng của điện áp bazơ-emitơ Vbe của tranzito trên và điện áp colectơ-emitơ Vce của tranzito dưới). Do đó, nếu dòng điện ngắn mạch được tạo ra là 10A, tổn hao 13W do nhiệt được

-2-

13839

tạo thành.

Do đó, ví dụ cần thiết có cánh tản nhiệt lớn để làm mát các tyristo S1 đến S3.

図 2.1-2 ベトナム語公報のイメージデータの一例（その 2）
（赤枠の中がノイズとなって残っている登録番号）

下のテキストは、上のイメージに対応する読み取り結果である。「-2-」というページ番号は削除されているが、赤枠で囲んだ「13839」という登録番号がテキストに含まれてしまっている。登録番号は文書を読み取る時点で明らかではあるが、上で述べたページ番号と違い機械的な削除処理はしていないと推察され、人手による読み取り結果の修正の際の削除漏れと考えられる。

Lượng nhiệt này là tích của dòng điện ngắn mạch được tạo ra và điện áp mở của các tyristo S1 đến S3. Ví dụ, điện áp mở của các tyristo S1 đến S3 khoảng 1,3 V (tổng của điện áp bazơ-emitơ Vbe của tranzito trên và điện áp colectơ-emitơ Vce của tranzito dưới). Do đó, nếu dòng điện ngắn mạch được tạo ra là 10A, tổn hao 13 W do nhiệt được 13839 tạo thành.

Do đó, ví dụ cần thiết có cánh tản nhiệt lớn để làm mát các tyristo S1 đến S3.

（３－２）文字認識誤りの傾向

DigiPat で公開されている OCR テキストの読み取り誤りの傾向として顕著なものは主に次の２ 点であった。

・声調記号の誤り

元のイメージデータでは声調記号がついている文字が、認識結果ではついていないという誤りが多い（例：誤[ê]→正[ế]）。この種の誤りを含むテキストを対訳コーパスとして統計翻訳(SMT)の学習を行う場合には、声調記号を削除して使用するということも考えられるが、これが有効かを判断するには、ある程度の規模の対訳コーパスを用いた実験による検証が必要である。

・数字の誤り

数字の１がＩ（アイ）や1(エル)などに誤って読み取られている場合が多い。

以上のような OCR の誤りには、認識候補文字が複数得られた場合に辞書や言語モデルを用いた後処理で解決できるものもあるため、今後ベトナム語の OCR でもエラー判定がさらに高精度化して正解率がさらに高くなる可能性も残っている。

２．１．２　文分割ツールの調査

（１）利用可能なツールの調査

ベトナム語用文分割ツールの調査対象として、トークナイザの前処理機能として文分割機能を有する vnTokenizer と、調査者が保有している英語特許文書の文分割用に開発した機械学習ベースの文分割ツール（調査対象としてはツールの主要部分である Stanford Classifier）の２つを調査した。ベトナム語の文分割ツールとして英語用のツールを評価対象に加えたのは、ベトナム語は英語と同じく通常文末にピリオドが付与されるためである。さらに、ベトナム語の特許文献では英語の特許文献で Figure を Fig.と記載するような省略語も多用されている。そのため、言語としては英語とベトナム語で異なるものの、そのような省略語の後での文分割判定を適切に行える可能性があると考えたからである[6]。

（１－１） vnTokenizer

・利用条件

[6] 今回使用した英語用の文分割ツールにおいて訓練に用いているデータ数は正事例が約 6,000、負事例が約 2,600 である。

GNU General Public License の下で利用できる。

・入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

https://github.com/phuonglh/vn.vitk

・利用環境の構築方法

Java で実装されているため、利用するには Java のインストールが必要である。

・使用方法

ツールを使用するためのシェルコマンド(vnTokenizer.sh または vnTokenizer.bat)が提供されているので、OS に応じたコマンドを使用する。コマンド起動時に、原文ファイルと出力ファイルの名前をパラメータで指定する。

例： vnTokenizer.sh -i input_file -o output_file

オプションパラメータで -sd を指定するとトークナイズの前に文分割を実行するので、これを指定して文分割を行う。

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

Java が必要であるが、無料で入手できる。

（１－２）Stanford Classifier

・利用条件

GNU General Public License (v2 or later)の下で利用できる。

・入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

http://nlp.stanford.edu/software/classifier.html

・利用環境の構築方法

Java で実装されているため、利用するには Java のインストールが必要である。

・使用方法

Stanford Classifier は訓練可能な汎用の分類器である。本ツールを文分割ツールとして利用する場合、分類の根拠となる訓練データで分類器を訓練した後、文分割される候補の個所を素性データとして入力して、文分割する／しないの２値に分類する。実行前にファイル 入出力の定義.prop の中で trainFile = 訓練データ と testFile = 文分割候補 でファイル名と学習する n-gram 数等を定義した後､以下のコマンドを実行する。分類結果は標準出力に出力される。

```
java -cp stanford-classifier.jar
edu.stanford.nlp.classify.ColumnDataClassifier ¥
  -prop 入出力の定義.prop -printClassifier HighWeight
```

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

Java が必要であるが、無料で入手できる。文分割の対象テキストから文分割される候補の個所を素性データとして抽出するプログラムと、Stanford Classifier による分類結果をもとに文分割する場所で文を改行するプログラムを作成する必要がある。

（2）評価

表 2.1-2 に文献別の評価結果を示す。表中、「未分割」とは文末と判定して文分割を行うべき個所で分割しなかったケース、「誤分割」とは文の途中であるため文分割してはいけない個所で分割されたケースを示す。表から明らかなように、Stanford Classifier を用いた文分割ツールは未分割が多いのに対し、vnTokenizer の場合には誤分割が多いことが分かる。

今回評価に使用した文の総数は 7,473 文であるので、それぞれのツールの正解率は以下のようになり、全体としては Stanford Classifier を用いた英語用の文分割ツールの方が分割精度が高かった。

Stanford Classifier： 1- (334+44)/7,473=94.9%

vnTokenizer： 1-(100+438)/7,473=92.8%

なお Stanford Classifier は機械学習ツールであるので、これを用いた文分割ツールの場合は文分割の正誤事例を追加して学習することで、さらに分割精度を向上させることができる。

表 2.1-2 文分割の評価（ベトナム語）

#	出願番号(VN)	文数	StanfordClassifier		vnTokenizer	
			未分割	誤分割	未分割	誤分割
1	1-2007-01812	317	12	2	1	9
2	1-2005-01787	214	8	0	3	2
3	1-2005-00402	274	28	1	4	18
4	1-2008-02431	144	1	0	3	2
5	1-2008-02720	238	3	0	4	5
6	1-2008-00617	404	11	0	0	8
7	1-2008-00182	145	9	0	1	6
8	1-2004-00924	248	4	1	21	13
9	1-2008-01287	232	9	1	10	11
10	1-2005-00564	132	3	2	4	8
11	1-2012-03600	202	11	3	3	7
12	1-2012-03893	254	6	0	1	2
13	1-2012-03659	208	10	0	3	15
14	1-2013-00360	83	1	0	6	4
15	1-2012-02596	288	4	1	4	8
16	1-2005-01458	314	25	0	13	4
17	1-2006-00845	252	11	1	0	7
18	1-2002-00889	243	9	1	3	6
19	1-2005-01263	364	13	5	1	66
20	1-2006-00339	232	19	1	0	5
21	1-2007-00963	406	13	8	0	50
22	1-2004-01292	149	8	0	1	1
23	1-2004-01208	432	23	0	3	5
24	1-2005-00810	177	3	1	0	4
25	1-2006-02057	356	25	16	1	138
26	1-2012-02960	277	15	0	0	9
27	1-2012-01489	304	22	0	3	12
28	1-2012-00180	221	8	0	0	1
29	1-2011-03350	198	10	0	5	7
30	1-2011-02026	165	10	0	2	5
計		7,473	334	44	100	438

（3）まとめ

ここでは、文分割ツールとして分割精度が高かった英語用文分割ツールで用いた

Stanford　Classifier について、改めて次の観点でまとめる。

・ツールの入手方法
・ツールの利用環境の構築方法、使用方法
・ツールのコスト及びツールと依存関係にある他のソフトウェアのコスト
・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）
・ツールの特性（処理誤りの傾向等）
・ツールの処理品質
・ツールに入力したデータ量と出力された辞書・コーパス等の量[7]
・処理に要する演算コスト（マシン性能と処理時間）
・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

（３－１）Stanford Classifier

・ツールの入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

http://nlp.stanford.edu/software/classifier.html

・ツールの利用環境の構築方法、使用方法

Java で実装されているため、利用するには Java のインストールが必要である。

Stanford Classifier は訓練可能な汎用の分類器である。文分割ツールとして利用する場合、本ツールは分類の根拠となる訓練データで分類器を訓練した後、文分割される候補の個所を素性データとして入力して、文分割する／しないの２値に分類する。実行前にファイル 入出力の定義.prop の中で trainFile = 訓練データ と testFile = 文分割候補 でファイル名と学習する n-gram 数等を定義した後、下記を実行する。分類結果は標準出力に出力される。

```
java -cp stanford-classifier.jar
edu.stanford.nlp.classify.ColumnDataClassifier ¥
  -prop 入出力の定義.prop -printClassifier HighWeight
```

・ツールのコスト及びツールと依存関係になる他のソフトのコスト

ツール本体ならびに動作に必要な Java はどちらも無料で入手できる。文分割の対象テキストから文分割される候補の個所を素性データとして抽出するプログラムと、文分割する場所で文を改行するプログラムを作成する必要がある。

・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）

ベトナム語は英語と同じピリオドを punctuation とすることから、本調査では英語用に作成した学習データを用いた。学習に用いたデータ数は以下の通りである。

正事例(正しい分割事例)　：約 6,000
負事例(間違った分割事例)：約 2,600

[7] 本項目は文分割ツールには該当しないと思われるため記載は省略する。

・ツールの特性（処理誤りの傾向等）

精度は学習データに依存するので、ツール特有の処理誤りというものは特にない。

・ツールの処理品質

約 95%の処理精度が出ていることから対訳コーパス作成用としては十分な精度が得られていると判断できる。

・処理に要する演算コスト（マシン性能と処理時間）

CPU が Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz×40、メモリが 396GB のマシンで 30 ファイル(計 7,473 文)の処理に 68.2 秒であった。

・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

学習データによりベトナム語に対応可能。

２．１．３　トークナイザの調査

（１）利用可能なツールの調査

ベトナム語のトークナイザとして、文分割ツールとしても評価した vnTokenizer に加え、JVNSegmenter というツールを調査した。

（１－１）vnTokenizer

・利用条件

GNU General Public License の下で利用できる。

・入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

https://github.com/phuonglh/vn.vitk

・利用環境の構築方法

Java で実装されているため、利用するには Java のインストールが必要である。

・使用方法

ツールを使用するためのシェルコマンド(vnTokenizer.sh または vnTokenizer.bat)が提供されているので、それを使用する。コマンド起動時に、原文ファイルと出力ファイルの名前をパラメータで指定する。

例：　vnTokenizer.sh -i input_file -o output_file

オプションパラメータで -sd を指定するとトークナイズの前に文分割を実行する。

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

Java が必要であるが、無料で入手できる。

（１－２）JVNSegmenter

・利用条件

GNU General Public License の下で利用できる。

・入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

http://jvnsegmenter.sourceforge.net/

・利用環境の構築方法

Java で実装されているため、利用するには Java のインストールが必要である。

・使用方法

以下のように、起動時にモデルが格納されているディレクトリと原文ファイル名（もしくは原文ファイルが存在するディレクトリ）を指定する。処理結果は、原文ファイルの名前の末尾に”.pro”を加えたファイルに出力される。

java -mx1024M -cp [classpath] jvntextpro.JVnTextProTest -modeldir [modelDir] [options…] -input [infile/indirectory] (-filetype (filetype))

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

Java が必要であるが、無料で入手できる。

（２）評価

トークナイザはコーパス作成やフレーズテーブル作成の過程で用いるツールである。ここでは、これらの過程で vnTokenizer と JVNSegmenter のどちらを用いるべきかを判断するため、明細書 1 ページ分(約 600 音節)のテキストを用いた小規模の性能比較を実施した。

評価対象テキストに対する処理結果において両者の出力に違いがあった 29 か所をベトナム語の専門家が分析したところ、vnTokenizer の誤りが 7 件であるのに対し、JVNSegmenter の誤りは 22 件であったため、精度の点から vnTokenizer がよいと判断した。

なお vnTokenizer を、特許庁の貸与データから抽出した 30 万文のベトナム語特許要約文を用いて動作確認を実施したところ、異常終了した。大規模なテキストを処理する場合には、このような異常終了に対する対策を考慮しておく必要がある。

（３）まとめ

評価した vnTokenizer の処理精度以外の情報をまとめる。

（３－１）vnTokenizer

・ツールの入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

https://github.com/phuonglh/vn.vitk

・ツールの利用環境の構築方法、使用方法

Java で実装されているため、利用するには Java のインストールが必要である。

ツールを使用するためのシェルコマンド(vnTokenizer.sh または vnTokenizer.bat)が提供されているので、それを使用する。コマンド起動時に、原文ファイルと出力ファイルの名前をパラメータで指定する。

例： vnTokenizer.sh -i input_file -o output_file

・ツールのコスト及びツールと依存関係になる他のソフトのコスト

ツール本体ならびに動作に必要な Java はどちらも無料で入手できる。

・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）

使用していない。

・ツールの特性（処理誤りの傾向等）

ツール起動時に辞書などのデータを読み込むため起動に時間がかかる。

・ツールの処理品質

JVNSegmenter よりも語分割の精度は高いと判断できたが、30 万文の特許要約文を用いて動作確認を実施したところ、異常終了した。大規模なテキストを処理する場合には、このような異常終了に対する対策を考慮しておく必要がある。また下に示すように一般的なトークナイザや形態素解析器と比べると処理時間が非常に長い。文書単位で処理すると、同じ程度の長さの文書であっても、比較的短時間で終わるものと、短時間で処理される場合より 2 桁以上時間がかかる場合があることが分かった。プログラムに何らかの不具合が存在する可能性がある。

・処理に要する演算コスト（マシン性能と処理時間）

CPU が Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz×40、メモリが 396GB のマシンで 7,169 文の処理に 26.3 分。

・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

ベトナム語用のツールであるため対応可能。

２．１．４　コーパス作成ツールの調査

（１）利用可能なツールの調査

本調査の開始時点において、コーパス作成ツール（文アライメントツール）の調査対象としては、内山らのツールと Gale らの手法を実装したツール[8]の２つが候補として挙がっていた。Gale らの手法は、文の単語数などの統計情報のみを元に処理を行うので、対訳辞書を用いずに対応付けを行うことができる。それに対し、内山らのツールは対訳辞書を用いた手法であり、同義関係も考慮した対応付けを行うことができる。大規模な対訳コーパス

[8] Gale らの手法の実装を明示しているツールとしては例えば hunalign がある。http://mokk.bme.hu/resources/hunalign/（2017 年 3 月末現在)。なお hunalign でも辞書を用いた処理も可能である。

を作成することを前提として考えた場合、ある程度の規模の対訳コーパスができた時点でそれを元にした統計翻訳システムや対訳辞書も利用できるようになるので、内山らの手法の方が最終的に良い結果が得られる可能性が高いと考えられる。これまでの特許庁の各種事業での利用実績も考慮し、内山らのツールに候補を絞って調査と評価を行うこととした。

（１－１）内山らのツール

・利用条件

国立研究開発法人情報通信研究機構（NICT）との利用契約が必要である。

・入手方法

契約締結後にNICTから提供される。

・利用環境の構築方法

本ツールは ruby 1.8 で動作するため、同言語がインストールされている必要がある。

・使用方法

(a)事前準備

本ツールは語の対応に基づいて対応付けを行うため、単語単位に分かれていない言語は分かち書きし、さらに、対訳辞書により語の対応付けを行ってから文の対応付けを行う。英語と日本語、中国語については、それぞれの分かち書きツールと中日・英日の辞書を備えているので、本ツールだけで文の対応付けができる。それ以外の言語の組については文分割・分かち書き・語の対応付けをユーザが行う必要があるが、インターネット上で提供されている機械翻訳サービスや既存の翻訳ソフトを用いてベトナム語テキストを日本語や英語に翻訳した結果を利用することができる。また、トークナイザを利用して語に分割した後、別途用意した対訳辞書を参照し、日本語または英語に置換するいわゆる単語翻訳を行った結果を用いることも可能である。

(b)文アライメント処理の実行

ベトナム語と日本語の文アライメントを作成する場合、jdir/ に日本語テキストを、vdir/に辞書引き等でベトナム語を英語に変換したテキストをセットする。下記のコマンドを実行すると jvdir/ に日本語テキストと英語テキストの文対応が作成される。

```
ruby alignje.rb jdir/ vdir/ jvdir/ >& log.txt
```

その後で、英語テキストを元のベトナム語文に置き換えて日本語とベトナム語文のアライメントを得る。

・ツールのコスト

有料（具体的な金額は利用契約による）

・ツールと依存関係になる他のソフトのコスト

内山らのツールを動作させるために必要なソフトは、同ツールに同梱されているので

コストが別途発生することはないが、上で述べたように事前準備の際に、ベトナム語テキストを英語や日本語に翻訳するためのツールは用意しなければならないため、そこで有償のソフトやサービスを利用すればその分のコストが発生する。

（２）評価

（２－１）対訳コーパスの作成手順

本調査で実施した対訳コーパスの作成手順は以下の通りである。(越日対訳コーパスを作成する場合を例に説明する。)

①越英機械翻訳を行って、ベトナム語テキストを英語に翻訳する。

②得られた英語テキストと、ベトナム語テキストに対応する日本語テキストを、内山らのツールを用いてアライメントを行い、英日対訳テキストを作成する。

③②で得られた英日対訳テキストにおける英語のテキストを、英語に翻訳する前のベトナム語のテキストに置き換える。

通常内山らのツールを用いた対訳コーパス作成においては、ツールが出力する文対応の確からしさを示すスコアが特定の閾値以上のもののみを採用してコーパスの精度を確保する。しかし、今回の調査においては、ツールが出力した対訳を全て人手で修正することとなっているため、スコアによる絞り込みは実施していない。

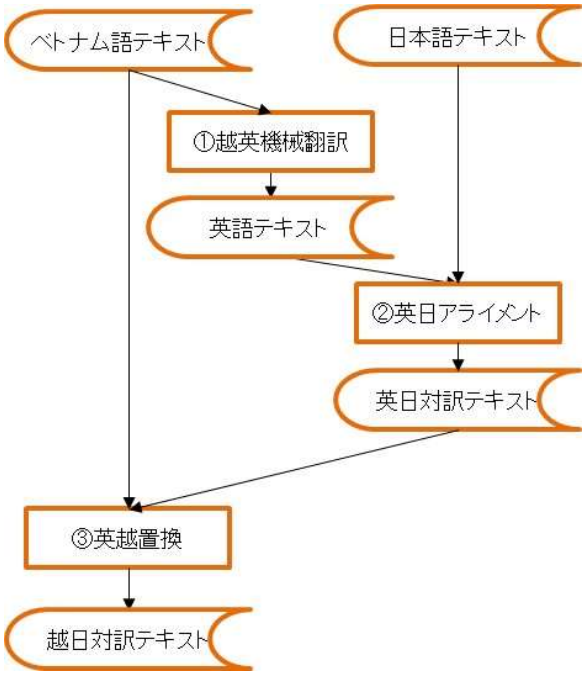


図 2.1-3　対訳コーパス作成手順の概要

なお、内山らのツールでは処理対象テキストの構文解析は行っていないので、①の翻訳

フェーズでの処理はベトナム語の個々の単語が英語に置き換わっているだけでよい。すなわち、翻訳においては文法的に正しい訳文を出力する必要はなく、対訳辞書があればそれを用いた辞書引き機能だけでベトナム語を英語に置き換えた形のいわゆる単語翻訳の結果を用いることも可能である。

（２－２）越日、越英対訳コーパスの作成

文分割ツールの処理結果に含まれる誤りを修正したベトナム語テキストと、それらを抽出したベトナム語明細書に対応する日本語ならびに英語明細書から抽出した日本語ならびに英語のテキストを用いて（２－１）で述べた手順により文アライメントを行った。その結果、以下の数のアライメント結果が出力された。

越日： 6,763 件

越英： 6,844 件

以下は出力された越日アライメント結果の一例である。第 4 文の日本語文に「///」というセパレータが入っているが、これは文が連結されたことを示しており、ベトナム語 1 文と日本語 3 文が対応していると推定された事例である。日本語の第 4 文はもともと「[0002] 従来から、クランク軸の一端に装着された駆動プーリ（プライマリーシーブ、又はドライブプーリ、とも云う。）と、変速軸の一端に装着された従動プーリ（セカンダリーシーブ、又はドリブンプーリとも云う。）をベルトによって連結してなる Ｖベルト式無段変速装置を内蔵してなる様々な Ｖベルト式無段変速装置内蔵エンジンが存在している。」という 1 文であったが、文アライメントツールの中では、一旦カッコ内の句点により 3 文に分割されており、それが対応付けの中で再び連結されたことが読み取れる。

#	score	ベトナム語	日本語
1	0.261807	Lĩnh vực kỹ thuật được đề cập	技術分野
2	0.079623	Sáng chế đề cập đến động cơ được lắp bộ truyền động biến thiên liên tục (CVT) nối liền trục khuỷu và trục truyền động mà chuyển động quay của trục khuỷu được truyền để làm thay đổi tốc độ giữa cả hai trục với chuyển động quay của trục khuỷu được truyền cho trục dẫn động.	[0001] 本発明は、クランク軸と、当該クランク軸の回転が伝達される変速軸とを連結して前記両軸間の変速を行う無段変速装置を内蔵し、前記変速軸の回転をドライブ軸に伝達する無段変速装置内蔵エンジンに関する。
3	0.130903	Tình trạng kỹ thuật của sáng chế	背景技術
4	0.088878	Các động cơ khác nhau có bộ truyền động biến thiên liên tục dạng hình chữ V được lắp cho động cơ thường có puli dẫn động (còn được gọi là puli sơ cấp) được lắp vào một đầu của trục khuỷu và puli bị dẫn (còn được gọi là puli thứ cấp) được lắp vào một đầu của trục truyền động được nối với nhau qua đai dạng hình chữ V.	[0002] 従来から、クランク軸の一端に装着された駆動プーリ（プライマリーシーブ、又はドライブプーリ、とも云う。///）と、変速軸の一端に装着された従動プーリ（セカンダリーシーブ、又はドリブンプーリとも云う。///）をベルトによって連結してなる Vベルト式無段変速装置を内蔵してなる様々な Vベルト式無段変速装置内蔵エンジンが存在している。

図 2.1-4 文アライメント結果の一例（越日）

以上のように作成した越日、越英対訳コーパスからそれぞれ 200 文をランダムに抽出し、本調査事業の仕様書に記載された以下の品質評価基準に従って評価を行った。

・評価基準

文対の文の内容が、ほぼ一致している（スコア A）

文対の文の内容が、50%以上一致している（スコア B）

それ以外（スコア C）

評価結果は以下の通りである。越英の場合スコア A が 92.0%なのに対し、越日では 72.0%しかなく、明らかに越日の精度が低いことが分かる。ベトナム語での明細書が英語版を翻訳する形で作成されていることが多いのが主要因と推察される。ちなみに、過去の特許庁の調査事業における中日機械翻訳用の辞書開発に関連する調査[9]においては、中国語文と日本語文が 1 対 1 で対応付けされ、かつそのアライメントのスコアが 0.08 以上のもののみを採用することで、文アライメントの正解率が 90%以上になるようにしている。

表 2.1-3 文アライメント結果の評価（越日、越英）

スコア	越日	越英
A	144（72.0%）	184（92.0%）
B	33（16.5%）	5（2.5%）
C	23（11.5%）	11（5.5%）

以下に各スコアに評価されたアライメント結果の一例を示す。表の#2 の例ではベトナム語のテキストに含まれる”Fig. 1”や”CVT”といったアルファベットのキーワードから推察されるように、ベトナム語テキストは英語テキストの後半と対応しているが赤字で示した前半部分に対応する部分がなく、スコア B と判断した[10]。また#3 の例では内容から判断して全く対応する部分がなくスコア C と判断した。

[9] 平成 25 年度「中国特許文献の機械翻訳のための新語に関する調査」

[10] 50%以上か否かの判断には厳密な文字数や単語数でのチェックはしておらず、あくまで評価者の主観によるものである。

#	ベトナム語	英語	スコア
1	Fig.3 là sơ đồ giải thích một ví dụ về qui trình sản xuất vật ghi đĩa quang theo một phương án của sáng chế	FIG. 3 explains an example of the process of producing the optical-disk recording medium as the embodiment of the present invention.	A
2	Fig.1 là hình chiếu cạnh trái của xe máy mà động cơ trên xe máy này được lắp CVT theo một phương án của sáng chế;	These and other features, aspects and advantages of the present invention will now be described with reference to the drawings of a preferred embodiment of the present invention, which embodiment is intended to illustrate and not to limit the invention, and in which figures: /// FIG. 1 is a left side view of a vehicle comprising an engine with an integrated CVT that has been arranged and configured in accordance with certain features, aspects and advantages of the present invention.	B
3	Trục truyền động 47 được lắp qua vòng bi 38 với nắp ngăn 71 ở bên phải đường tâm L của thân xi lanh 19 cũng như được lắp qua vòng bi 39 vào mặt đầu trái của ngăn thứ hai 41 ở bên trái.	Moreover, the illustrated configuration maintains the serviceability of the clutch. /// The right end of the transmission shaft 47 preferably extends into the transmission case 45 beyond the second case portion 41.	C

図 2.1-5　文アライメント結果（越英）の評価の一例

その後、文アライメント結果を全て人手でチェックし、最終的に以下の文対の対訳コーパスが得られた。

越日：　6,485

越英：　7,169

（３）考察

（３－１）単語翻訳による翻訳結果を用いた文アライメント

（２－１）で述べたように本調査においてはベトナム語テキストを最初に越英機械翻訳を行って英語のテキストにし、それと日本語テキストを対応づけることで、最終的にベトナム語と日本語のテキストを対応づけるというアプローチを取った。その際、越英機械翻訳にはインターネット上の翻訳サービスを利用したが、単純な辞書引きによる単語翻訳の結果を利用して同程度の文アライメント精度が得られるのであれば、大規模な対訳コーパスを作成する際には一つの選択肢となりうる。そこで、その可能性を検証するため、約 2,700 語の小規模な対訳辞書を用いて文アライメントを実施した[11]。検証に用いたのは文献#1 で、（２）における越日対訳コーパス作成の結果、300 の対訳が得られた文献である。

まず、対応関係に変化があった事例を分類した。表 2.1-4 がその結果である。表に示す通り 15 件(全 300 件の 5.0%)の対応関係に変化があり、60%はベトナム語側の文のみが変化したケースであった。

[11] ３．４．２で述べるように、本検討での利用を念頭に Wikipedia の見出しを用いて対訳辞書を試作したが、単語翻訳の結果を見ると、非技術系の見出しが多く文アライメントで精度向上に寄与する可能性が低いと思われた。そのためここでは、２．１．５（２－２）で述べる発明の名称を元に作成した対訳コーパスに出現する単語をもとに約 2,700 語の小規模な見出しリストを作成し、それにインターネット上の翻訳サービスで訳を付与して作成した対訳辞書を用いた。

表 2.1-4 文アライメント結果の差

種別	件数
ベトナム語の文のみが変化	9
日本語の文のみが変化	4
両方が変化	2
計	15

以下に、変化の種別の事例としてベトナム語の文のみが変化した場合と、日本語の文のみが変化した場合の事例を示す。各事例において赤字で示した文が変化した文である。

Cần lưu ý là, trên thực tế, mã phát hiện lỗi và mã sửa lỗi được gắn vào dữ liệu người sử dụng, và dữ liệu người sử dụng được xử lý đan xen và xử lý khác. /// Điều biến độ dài thay đổi được thực hiện ở bước S12.	なお、実際には、ユーザデータに対する誤り検出符号及び誤り訂正符号の付加、インターリーブ処理等も行われる。
Dãy dữ liệu thu được từ việc tạo khuôn ở bước S11 được điều biến độ dài thay đổi.	可変長変調工程 S12では、フォーマツトイヒ工程 S11により生成されたデータ列に対して可変長変調処理を施す。

(a)機械翻訳を用いた文アライメント結果

Cần lưu ý là, trên thực tế, mã phát hiện lỗi và mã sửa lỗi được gắn vào dữ liệu người sử dụng, và dữ liệu người sử dụng được xử lý đan xen và xử lý khác.	なお、実際には、ユーザデータに対する誤り検出符号及び誤り訂正符号の付加、インターリーブ処理等も行われる。
Điều biến độ dài thay đổi được thực hiện ở bước S12. /// Dãy dữ liệu thu được từ việc tạo khuôn ở bước S11 được điều biến độ dài thay đổi.	可変長変調工程 S12では、フォーマツトイヒ工程 S11により生成されたデータ列に対して可変長変調処理を施す。

(b)単語翻訳を用いた文アライメント結果

図 2.1-6 機械翻訳と単語翻訳を用いた文アライメントの変化(1)
(ベトナム語の文のみが変化した事例)

Một RUB bao gồm 16 đơn vị địa chỉ ("cung" như được thể hiện) và hai khung liên kết.	1つの RU Bは、 16個のアドレスユニット（図中「Sector」で示す。 ///）と、 2つのリンキングフレームから構成される。
Từng khung liên kết được bố trí như là vùng đệm giữa các RUB.	リンキングフレームは、各 RUB間の緩衝領域として設けられている。

(a)機械翻訳を用いた文アライメント結果

Một RUB bao gồm 16 đơn vị địa chỉ ("cung" như được thể hiện) và hai khung liên kết.	1つの RU Bは、 16個のアドレスユニット（図中「Sector」で示す。
Từng khung liên kết được bố trí như là vùng đệm giữa các RUB.	）と、 2つのリンキングフレームから構成される。 /// リンキングフレームは、各 RUB間の緩衝領域として設けられている。

(b)単語翻訳を用いた文アライメント結果

図 2.1-7 機械翻訳と単語翻訳を用いた文アライメントの変化(2)
（日本語の文のみが変化した事例）

次に、文アライメントの修正結果である正解データと比較し、これらの変化が改善であるか悪化であるかを判定した。例えば、上で示した事例においてベトナム語の文のみが変化した事例(図 2.1-6)は改善した事例、日本語の文のみが変化した事例(図 2.1-7)は悪化した事例である。

下の表に評価結果を示す。表から明らかなとおり、越英機械翻訳を使って翻訳した結果を用いて文アライメントを行った場合と比較して悪化している事例の方が多いものの改善している事例もあり、今回用いた対訳辞書よりもさらに規模の大きな辞書を用いれば、単語翻訳でも文アライメントを十分行うことができる可能性があると結論付けられる。

表 2.1-5 文アライメント結果の変化の評価

種別	件数
改善	5
悪化	8
同等	2
計	15

（３－２）文アライメントツールのスコアと評価のスコアについて

本調査においては文アライメントの結果を全て人手でチェックし、不適切な対応づけがあった場合にはそれらを全て修正して対訳コーパスを作っているが、対訳事例の数が 100 万文対を超えるような大規模な対訳コーパスを作成する場合には、今回のように文アライメント結果を全てチェックするのは現実的ではない。そのため、通常は文アライメントツールが処理結果の一部として出力する、対応の度合いを示すスコアや対応する文の数の情報を基にして絞り込みを行い、絞り込み条件に合致した対訳事例は人手チェックをせずに

そのまま利用するという形を取ることが多い。（２－２）でも述べた通り、中日機械翻訳のための辞書整備に関する特許庁の過去の事業においても、文アライメントツールのスコアが 0.08 以上で文の対応は 1 対 1 のものに限定する条件で文アライメント結果を絞り込んで利用している[12]。今後ベトナム語の対訳コーパスを大規模に整備していく場合には、中日の場合と同様に、各言語対に応じて適切な絞り込み条件を設定することが重要となるが、ここではその最初の足掛かりとして、文アライメントの評価で使用した 200 事例の評価結果を分析し、大まかな傾向を把握することにする。

図 2.1-8 は文アライメントツールのスコアに対して、どのような品質評価基準のスコアが付与されたかを示したものである。それぞれの言語対での文アライメント結果をアライメントツールのスコアが 0.4 以上、0.3 以上 0.4 未満、0.2 以上 0.3 未満、0.1 以上 0.2 未満、0.1 未満の 5 つにグループ分けし、人手による評価のスコア A,B,C がそれぞれ何文対あるかを示している。これらのグラフから(b)越英に関しては上で紹介した中日の事例と同様、スコア 0.1 付近で絞り込める可能性が読み取れるが、(a)越日ではスコア 0.1 以上であっても B や C 評価の文対応が含まれてしまい、仮に B や C 評価のものを振るい落とすためには絞り込みの閾値を高くすると、今度は本来抽出したい正しい対応の事例も採用できなくなる可能性が高くなることが分かる。

[12] この絞り込み条件は、平成 25 年度「中国特許文献の機械翻訳のための新語に関する調査」や平成 26 年度「中国特許文献の機械翻訳のための辞書整備及び機械翻訳の品質評価に関する調査」等で利用されている。

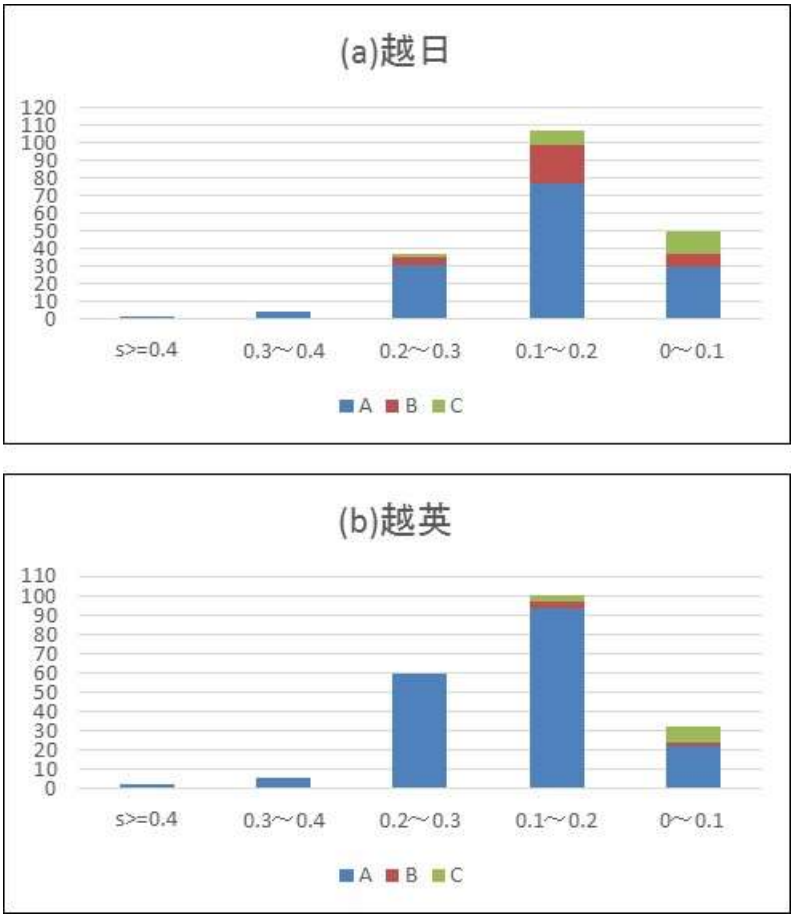


図 2.1-8 文アライメントスコアに対する文対応の評価結果(1)

そこで次に、スコア A の割合が文アライメントのスコアに応じてどのように変化するかを調べた。図 2.1-9 がその結果である。(b)越英では、文アライメントのスコアが 0.10 以上 0.12 未満のものにおけるスコア A の割合が 0.89 であり、上述した中日での事例と似通った結果となった。一方、(a)越日に関しては、アライメントのスコアが 0.18 以上 0.20 未満の場合でもスコア A の割合は 80%を切る場合もあり、0.16 未満でのスコア A の割合は 60～70%にすぎない。これは、従来の中日の場合や、２．２．４で述べる泰日や泰英の場合と比べて大きく異なっている傾向であり、今後アライメント精度が低い原因についてさらに分析を進める必要がある。

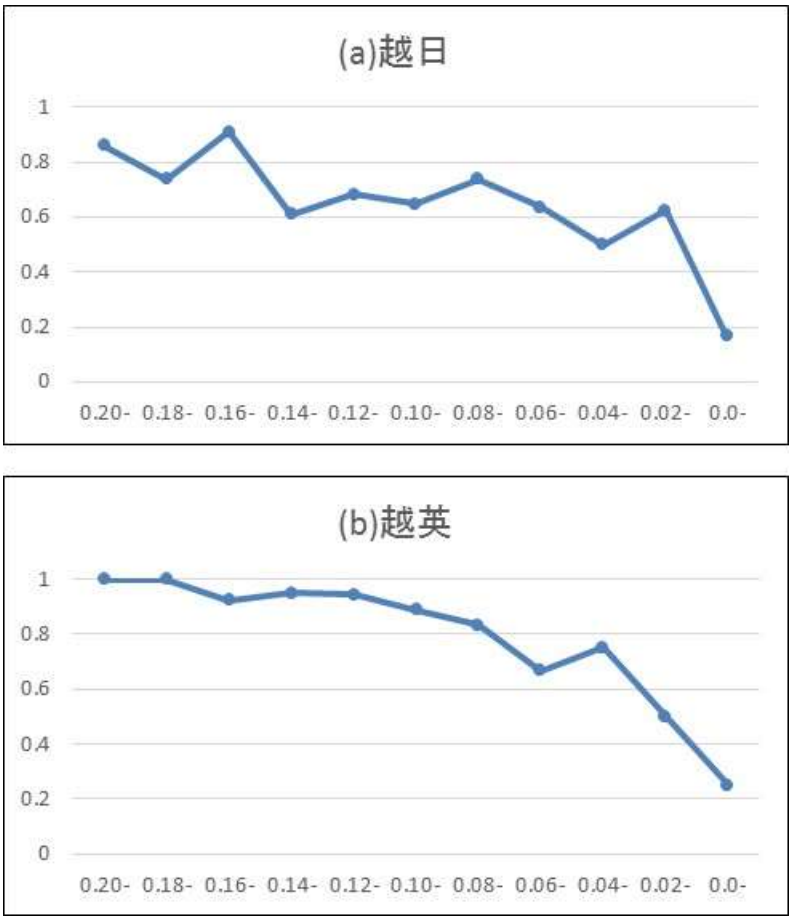


図 2.1-9 文アライメントスコアに対する文対応の評価結果(2)

続いて下表は、文アライメントにおいて、何文と何文が 1 つの対訳事例として抽出されたのか、そしてそれらがどのように評価されたかをまとめたものである。例えば、(a)越日では、ベトナム語 1 文と日本語 1 文からなる文対が、評価した 200 例のうち 93.0%の 186 件あり、それらのうち 139 件(186 件の 74.7%)が A 評価であったことを示している。また、(b)越英では、1:1 の文対が 190 件(95.0%)あり、A 評価は 179 件(94.2%)である。一方 1:1 以外の事例での A 評価は、越日が 14 件中 5 件(36%)、越英が 10 件中 5 件(50%)と正解率が低く、1 対 1 対応以外の対訳事例を用いるには人手によるチェックが不可欠であると言える。

表 2.1-6 対応づけされた文数別のアライメント評価結果

(a)越日

	A	B	C	計
1:1	139	28	19	186
1:5	1	0	0	1
2:1	4	4	1	9
3:1	0	1	0	1
4:1	0	0	2	2
5:1	0	0	1	1

(b)越英

	A	B	C	計
1:1	179	4	7	190
1:2	1	0	0	1
2:1	3	1	0	4
2:3	1	0	0	1
4:1	0	0	1	1
5:1	0	0	2	2
5:2	0	0	1	1

（３－３）原出願の違いによる文アライメントの傾向

出願人が日本企業か米国企業かにより、ベトナム語に翻訳するプロセスが異なる可能性があり、それが文アライメントや最終的に辞書登録候補語の抽出精度に関係してくる可能性がある。すなわち、出願人が米国企業の場合は、通常、英語の原出願から直接ベトナム語と日本語に翻訳すると思われるため、日越と比べて英越の方が文対応が取りやすいと考えられる。一方、出願人が日本企業の場合は、ベトナム語に翻訳するには、日本語から直接ベトナム語に翻訳する場合と、先に日本語を英語に翻訳してからそれをさらにベトナム語に翻訳する場合があると考えられる。後者のように一旦英語にすると、翻訳を二回重ねることになるため、翻訳過程を重ねるごとに原文との間での単語対応が取れない語が訳出されたり翻訳誤りが入ったりする可能性が高くなり、結果としてそのように英語を経由して日本語からベトナム語に翻訳された明細書は日本語から直接翻訳された場合に比べ、文アライメントや辞書登録候補語の抽出の精度が低い可能性がある。

この可能性を検証するため、原出願が日本か米国かで文アライメントの精度にどのような違いがあるかを各言語対の文アライメントの評価で使用した 200 文の評価結果を用いて分析した。図 2.1-10 がその結果である。図から明らかなとおり、(b)越英の場合は、原出願の差はほとんどないが、(a)越日の場合には、明らかに日本が原出願の場合の文アライメント精度が悪いことが分かる。

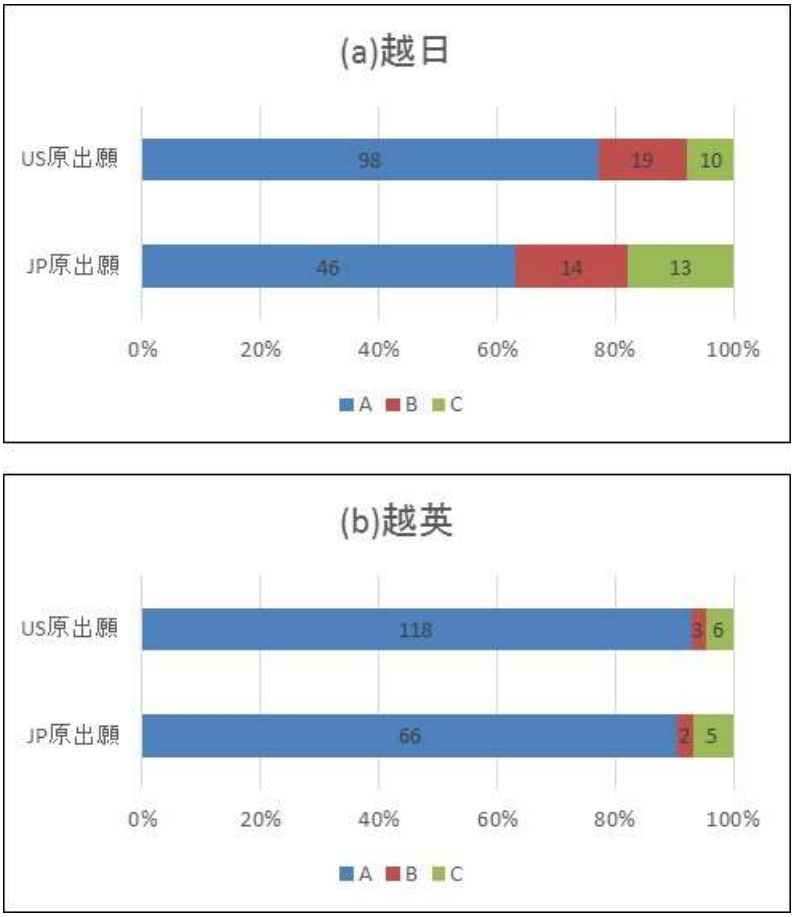


図 2.1-10 原出願国別の文アライメントの評価（越↔日／英）

（4）まとめ

ここではコーパス作成ツール（文アライメントツール）の試用結果を、次の観点でまとめる。

・ツールの入手方法
・ツールの利用環境の構築方法、使用方法
・ツールのコスト及びツールと依存関係にある他のソフトウェアのコスト
・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）
・ツールの特性（処理誤りの傾向等）
・ツールの処理品質
・ツールに入力したデータ量と出力された辞書・コーパス等の量
・処理に要する演算コスト（マシン性能と処理時間）
・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

（4－1）内山らのツール

・ツールの入手方法

利用にあたってはNICTとのライセンス契約が必要であり、契約締結後にNICTから提供される。

・ツールの利用環境の構築方法、使用方法

本ツールは ruby 1.8 で動作するため、同言語がインストールされている必要がある。使用法については、（１）を参照のこと。

・ツールのコスト及びツールと依存関係になる他のソフトのコスト

有料（具体的な金額は利用契約による）。ツールを動作させるために必要なソフトは、同ツールに同梱されているのでコストが別途発生することはない。

・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）

なし。

・ツールの特性（処理誤りの傾向等）

対応づけを行う２つの文書で、文の出現順序を変えることなく対応付けができるという前提で処理を行うため、対応する文が存在しない場合でも、その文の前後も含めて誤った対応付けをしてしまう場合がある。したがって、国によってクレームと詳細な説明の出現順が違う特許文書の対応付けを行うような場合には、それらを予め定めた順序に変更することや、クレームとそれ以外を別の文書として扱うなどの利用上の工夫が必要である。

・ツールの処理品質

特許庁においても英日、中日対訳コーパスの開発に使用されており、処理品質が問題になることはないと思われる。

・ツールに入力したデータ量と出力された辞書・コーパス等の量

ツールに入力したデータと出力されたアライメントの数は以下の通りである[13]。

表 2.1-7 ツールに入力したデータ量と出力されたアライメント数

項目		越日	越英
入力	ベトナム語文数	7,473	7,473
	日本語文数	7,809	－
	英語文数	－	7,560
出力		6,763	6,844

・処理に要する演算コスト（マシン性能と処理時間）

越日アライメント(30 文献)の処理を行うケースで、RedHat 上のバーチャルマシン(ホストマシンのスペック： CPU が Intel(R) Xeon(R) E5-2690 v2 @ 3.00GHz×20、メモリが 32GB)にて 13.1 分。

・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

[13] ベトナム語の文数は、文分割ツールの処理結果を人手で修正した後の文の数であるが、日本語と英語の文の数は本ツールの文分割機能によって分割された文の数であり、誤りを含んでいる可能性がある。そのため、実際の文の数はこれらとは異なる可能性がある。

調査者が NICT よりライセンスを受けている内山らのツールは、英日・中日対訳コーパスを開発するためのツールとしてパッケージングされたものであり、本調査にあたっては越日コーパスを作るために一旦ベトナム語のテキストを英語に翻訳してから使うといったイレギュラーな使い方をしている。

２．１．５　辞書作成ツールの調査

（１）利用可能なツールの調査

辞書作成ツール（フレーズテーブル作成ツール）としては、統計翻訳用のツールキットである Moses が研究目的ならびに実用目的で広く使用されていることから、当該ツールと、そのツールと併せて使用する単語アライメントツールである mgiza に対象を絞って調査・評価を行った。

（１－１）Moses

・利用条件

GNU Lesser General Public License(LGPL)の下で利用できる。

・入手方法

以下の URL からソースコードをダウンロードする(2017 年 3 月末現在)。

https://github.com/moses-smt/mosesdecoder

・利用環境の構築方法

Moses のビルドに必要な各種パッケージを事前にインストールする必要がある。詳細は、以下の URL で公開されているマニュアルに記載されている。

http://www.statmt.org/moses/manual/manual.pdf

・使用方法

1)対訳コーパスの原言語、目的言語の各テキストをトークナイザで単語もしくは形態素列に分割する。

2) clean-corpus-n.perl を用いて(1)で作成したデータをクリーニングする。

3) train-model.perl を用いて単語アラインメントを実施する。起動に際して各種パラメータの設定が必要であるが、詳細は利用環境の構築方法で述べたマニュアルに記載されている。

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

ツールの実行に必要な種々のソフト(例えば Boost ライブラリや言語モデル用ツール irstlm 等)が存在するが、それらはいずれも無料で利用できる。

（１－２）mgiza

・利用条件

GNU General Public License の下で利用できる。

・入手方法

以下の URL からソースコードをダウンロードする(2017 年 3 月末現在)。

https://github.com/moses-smt/mgiza

・利用環境の構築方法

mgiza をコンパイルするには、Boost ライブラリを事前に使える状態にしておく必要がある。

ソースコードをダウンロード後、展開したソースディレクトリで以下のようにビルドする。

```
cmake .
make
make install
```

・使用方法

Moses で提供される訓練ツールである train-model.perl において、起動時のパラメータに　-mgiza　を加えることで使用する。併せて、並列処理のために使用する CPU の数を -mgiza-cpus NUMBER で指定する。

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

mgiza をコンパイルするのに必要な Boost ライブラリは無料で利用できる。

（２）評価

（２－１）辞書登録候補語の作成手順

辞書登録候補語の作成手順は以下の通りである。(日越辞書を作成する場合を例に説明する。)

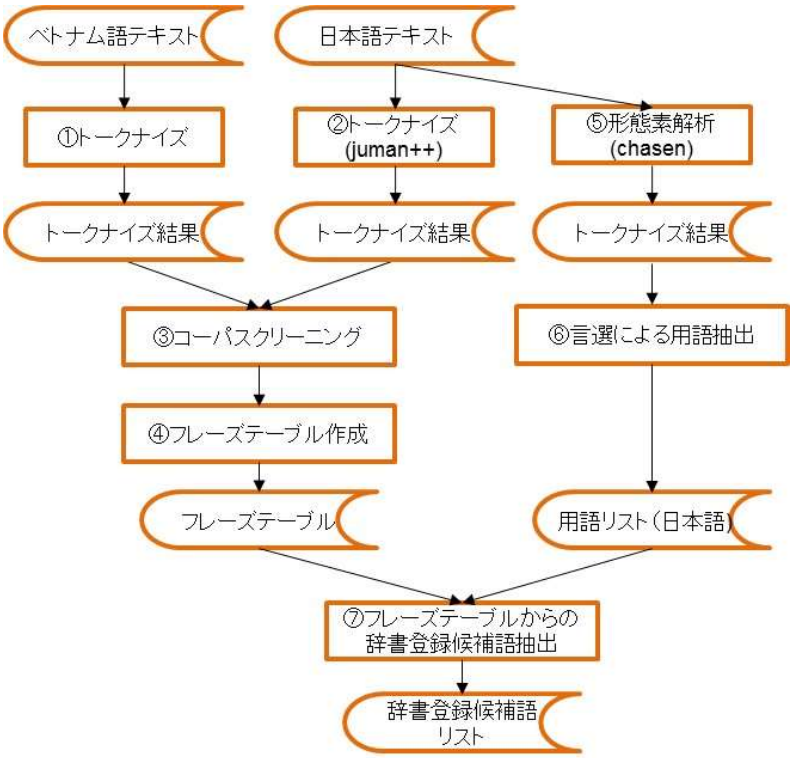


図 2.1-11 辞書登録候補語の作成手順の概要

（２－１－１）Moses を用いたフレーズテーブルの作成

Moses を用いたフレーズテーブルの作成は以下の手順で行った。

①ベトナム語の訓練データを vnTokenizer（タイ語の場合は OpenNLP）を用いてトークナイズする。

②日本語の訓練データを juman++を用いてトークナイズする。

③トークナイズ結果に対して clean-corpus-n.perl を用いてクリーニングを行う。

④train-model.perl を用いて越日統計翻訳の学習を行い、フレーズテーブルを作成する。

なお、訓練データには、２．１．４で作成した対訳コーパスに加え、下の（２－２）で述べるパテントファミリーを用いて作成した発明の名称の対訳コーパスと、３．４．２で述べる Wikipedia の見出しから作成した越日対訳辞書も利用した。

（２－１－２）言選を用いた辞書登録候補語の抽出

言選は、日本語および英語のテキストから、専門用語（キーワード）を抽出するツールである[14]。ツールに対する入力データは、同ツールが指定する形態素解析ツールの形式で用意する必要がある。

⑤日本語テキスト[15]を Chasen[16]を用いて形態素解析する。（英語の場合は Brill tagger に

[14] http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html （最終検索日:2017 年 2 月 20 日）

[15] 用語の抽出には２．１．４で作成した対訳コーパスの日本語テキストを用いた。パテン

よる解析結果が必要である。本調査では Brill tagger と同じ出力形式で解析結果を出力できる GPoSTTL[17]を用いた。）

⑥言選を実行して用語リストを作成する。

⑦（１－１）で作成したフレーズテーブルから⑥で作成した用語リストの見出しに合致するフレーズを、翻訳確率の最も高いものから上位３件抽出する[18]。なお、フレーズテーブルのフレーズは、通常の辞書の見出しと違い、助詞や句読点などを含んだ見出しとなっている場合があるため、フレーズテーブルからのフレーズの抽出にあたってはそれらの存在を考慮した処理が必要である[19]。

（２－２）学習に用いる発明の名称の対訳コーパスの作成

今回の評価で対訳コーパスの試作に用いる明細書の数は、越英・越日対訳の場合で 30 文書であり、対訳コーパスのサイズで見ても越日が 6,485 文対、越英が 7,169 文対とそれほど多くない。通常の統計翻訳において学習に用いる対訳コーパスのサイズと比べて極めて小さいと考えられる。そこで、特許ドメインでの対訳コーパス量を増やすため、パテントファミリーの発明の名称を抽出して対訳コーパスを作成した。

ベトナム語の特許に対応する日本語ならびに英語のパテントファミリーの発明の名称は、ベトナム国家知的財産庁が提供する登録特許用のデータベース DigiPat を利用して以下のように抽出した。

①フロントページ検索で検索キーワードとして「PCT JP」を指定して、優先権主張番号に JP（英語用は US）を含むものを検索する。

トファミリーを用いて作成した発明の名称の対訳コーパスと、Wikipedia から作成した越日対訳辞書は利用していない。

[16] 形態素解析システム茶筌 http://chasen.naist.jp/hiki/ChaSen/ （最終検索日：2017 年 3 月 17 日）

[17] http://gposttl.sourceforge.net/（最終検索日：2017 年 2 月 20 日） Brill tagger の出力形式で出力するためにはコマンド起動時の引数として--brill-mode を指定する必要がある。

[18] 学習に用いる訓練データの規模が大きい場合には翻訳確率が一定数以下のものは抽出しないということも考えられるが、今回はそのような条件は設けずフレーズテーブルに登録されているものは全て抽出対象とした。

[19] 辞書登録候補語をフレーズテーブルから抽出する際、既存の日英辞書の日本語見出し、英語見出しと一致するフレーズを優先することで候補作成精度を向上させる可能性も検討した。フレーズテーブルの見出しはトークナイザが分割した形態素に分かれているため、日本語見出しに不要な助詞が含まれていても簡単に削除でき、既存の辞書の日本語見出しと比較しなくても、名詞であれば高い精度で抽出できる。よって、上記のような既存辞書の見出しとの比較は不要であると判断した。ただし、蓄積語数が多くなって、例外的な語（例えば、「は虫類」「しきい値」等の平仮名を含む名詞など）の採取も必要になった場合には有効である可能性がある。

図 2.1-12 DigiPat のフロントページ検索画面

②検索結果のソースを手動もしくは機械的に HTML ファイルとして保存する。

図 2.1-13 DigiPat での検索結果の表示例

③保存した HTML ファイルからベトナム語のタイトルを取り出す。

④同じく保存した HTML ファイルから優先権主張番号を取り出し、それをもとに日本語明細書を検索し日本語のタイトルを取得する。(英語の場合も同様)

以上の手順により、以下の件数のパテントファミリーが得られ、各言語での発明の名称を用いて同件数の対訳コーパスが得られた[20]。

越日： 990 件

越英： 1,767 件

（２－３）日越、英越辞書の作成と評価

（２－１）で述べた手順に従い日越、英越辞書を作成した。作成の過程で得られたデータ数、ならびに最終的に得られた辞書登録候補語の数は以下の通りである。

[20] これらの件数は発明の名称が同一のものを含んだ数字である。

表 2.1-8 日越、英越辞書作成に関連するデータ数

		日越	英越
訓練データ数(クリーニング前)		170,883	284,882
内訳	明細書対訳	6,485	7,169
	名称対訳	990	1,767
	対訳辞書[21]	163,408	275,946
訓練データ数(クリーニング後)		170,234	284,199
フレーズ数 (フレーズテーブルのサイズ)		754,075	1,150,719
言選が抽出した用語数		11,220	14,873
フレーズテーブルから抽出できた 辞書登録候補語の数		2,250	6,457

得られた辞書登録候補語の一例として日越辞書の候補語を示す。

表 2.1-9 日越辞書登録候補語の例

#	日本語	ベトナム語 1	ベトナム語 2	ベトナム語 3
1	繊維	từ sợi	các sợi	đều sợi
2	表面	phía bềmặt	bềmặt bằng	được phủ bởi bềmặt
3	データ	dữliệu	dữliệu sẽ	để tập dữliệu
4	シリカ	silic dioxit	bằng silic oxit	đốivới silic oxit
5	量	đạilượng	lượng nằm	đốivới lượng
6	図	các hìnhvẽ từ Fig	trên FIG	ống nano thểhiện
7	重量	theo khốilượng	khốilượng	xácđịnh trọnglượng
8	データレート	Tốcđộ dữliệu	tốcđộ dữliệu	và tốcđộ
9	発明	Phátminh	Sángchế	ưutiên của
10	酸	Axit	axit được	và axit
11	値	trịsố	giátrị	đọc có giátrị
12	化	hóa được	hóa	· hóa
13	水	Nước	nước	đồng sôi
14	樹脂	nhựa	nhựa và	với nhựa
15	炭素	cacbon	Cacbon	than
16	方向	chiều	dọc	đổi hướng
17	式	Kiểu	đặcsắc	có côngthức
18	温度	nhiệtđộ là	nhiệtđộ	ở nhiệtđộ
19	実施	trong Vídụ	thểhiện theo các	ưutiên
20	物質	vậtchất	ở đó chất	từ đó chất

[21] ここで用いた対訳辞書は３．４．２で作成した Wikipedia の見出しを用いて試作した対訳辞書である。

このように作成した候補語リストの先頭 200 語を抽出し、本調査事業の仕様書に記載された以下の品質評価基準に従って評価を行った。

辞書登録候補語の評価基準：

両言語のフレーズ・語の意味が、おおむね一致している(スコア A)

両言語のフレーズ・語の意味が、一部重なるものである(スコア B)

両言語のフレーズ・語の意味は、ほとんど共通しないものである(スコア C)

図 2.1-14 及び図 2.1-15 がその結果である。第 1 訳語だけを見ても、スコア A と判定された割合が日越は 86.5%、英越が 67.5%である。この数値は中日機械翻訳のための辞書開発事業において行った大規模な辞書開発での抽出精度の同等以上の数字となっており、この精度で大規模な抽出ができれば、処理精度としては実用レベルであると言える[22]。

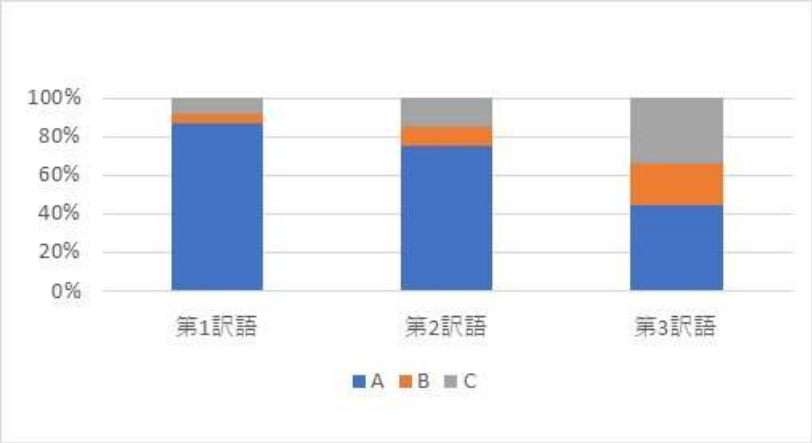


図 2.1-14 日越辞書精度

[22] 平成 25 年度「中国特許文献の機械翻訳のための新語に関する調査」では 100 万語規模で実施された中日対訳辞書候補抽出の精度が約 70%であったと報告されている。

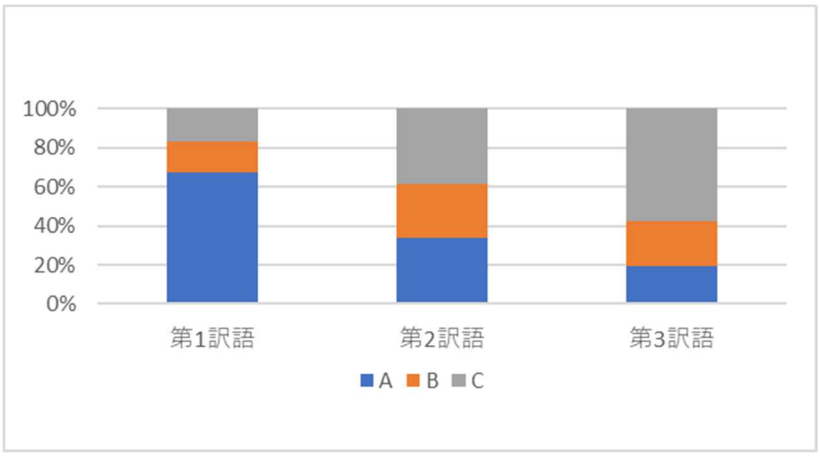


図 2.1-15 英越辞書精度

（３）考察

（３－１）辞書登録候補語の抽出に使用するフレーズテーブルについて

日越と越日フレーズテーブルを作成して、ある日本語とベトナム語のペアが越日辞書候補と日越辞書候補の両方に存在している場合は、辞書候補として正しい可能性が高いと考え、この仮説を検証した。

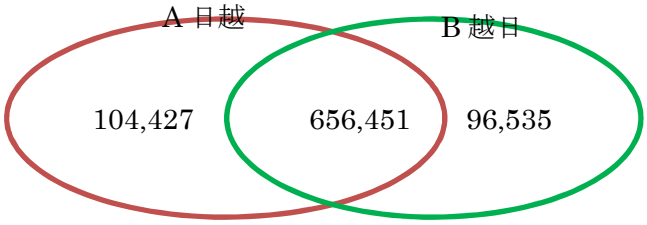
結果としては、両方に存在する候補を優先した場合でも正解率はほとんど変化しなかった。

（３－１－１）フレーズテーブルの分析

（２－３）で用いた越日フレーズテーブル(データ数 754,075)に加え、日越方向での学習を行って日越フレーズテーブル（データ数 761,977）を作成し、それらのフレーズテーブルのベトナム語と日本語の組の重なりを調査した。なお、日本語またはベトナム語の一方が空白等の明らかに無意味なデータを除外してカウントした。その結果以下のようになった。全体の 76.5%(656,451/857,413)は二つのフレーズテーブルに存在するフレーズの組であるが、23.5%は異なるフレーズの組である。

表 2.1-10 日越・越日フレーズテーブルの重なり

分類	データ数
ベトナム語と日本語の組が A 日越だけに存在	104, 427
ベトナム語と日本語の組が AB 日越と越日両方に存在	656, 451
ベトナム語と日本語の組が B 越日だけに存在	96, 602
計	857, 413



（３－１－２） 評価

日越方向のフレーズテーブルから辞書登録候補語を抽出したところ、（２－３）において作成した越日方向のフレーズテーブルから抽出した辞書登録候補語数とほぼ同数の 2, 276 語が抽出できた。表 2.1-11 に、抽出結果の一部を（２－３）で抽出済の辞書登録候補語と対比する形で示す。赤字で示したベトナム語が、今回日越方向のフレーズテーブルから抽出したベトナム語と、（２－３）で抽出した越日方向のフレーズテーブルから抽出したベトナム語で共通するものである。

表 2.1-11 辞書登録候補語の例

#	日本語	学習方向	ベトナム語 1	ベトナム語 2	ベトナム語 3
1	繊維	越日	từ sợi	các sợi	đều sợi
		日越	các sợi	sợi	đều sợi
2	表面	越日	phía bề mặt	bề mặt bằng	được phủ bởi bề mặt
		日越	Surface Systems	các bề mặt	đế thủy tinh có các bề mặt
16	方向	越日	chiều	dọc	đối hướng
		日越	hướng	dọc	đối hướng

（２－３）の評価では、例えば#1 のベトナム語 1「từ sợi」はスコア B(両言語のフレーズ・語の意味が一部重なる)と評価された。一方第 2 訳語として抽出された「các sợi」はスコア A と評価されたが、今回の抽出では第 1 訳語となっており、このように両方に存在する

訳語を優先することで第 1 訳語の精度を向上できる可能性がある。

この可能性を検証するため、（２－３）の評価で用いた 200 語に対し、上で述べたように共通するベトナム語がある場合にそれを第 1 訳語として採用した場合のスコアの変化を調べた。その結果、下表に示すように 200 語中 21 語でスコアの変化があったものの、改善と悪化の事例がほぼ同数であり、期待した効果は得られなかった。

表 2.1-12 第 1 訳語の評価の変化

スコアに変化があった訳語	21
改善	10
悪化	11

（４）まとめ

ここではフレーズテーブル作成に関連するツールの試用結果を、次の観点でまとめる。

・ツールの入手方法
・ツールの利用環境の構築方法、使用方法
・ツールのコスト及びツールと依存関係にある他のソフトウェアのコスト
・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）
・ツールの特性（処理誤りの傾向等）
・ツールの処理品質
・ツールに入力したデータ量と出力された辞書・コーパス等の量
・処理に要する演算コスト（マシン性能と処理時間）
・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

（４－１）Moses ならびに mgiza

・ツールの入手方法

それぞれ以下の URL にてダウロード可能(2017 年 3 月末現在)。

Moses: https://github.com/moses-smt/mosesdecoder

mgiza: https://github.com/moses-smt/mgiza

・ツールの利用環境の構築方法、使用方法

- 構築方法

Moses については、ビルドに必要な各種パッケージを事前にインストールする必要がある。詳細は、以下の URL で公開されているマニュアルに記載されている。

http://www.statmt.org/moses/manual/manual.pdf

mgiza をコンパイルするには、Boost ライブラリを事前に使える状態にしておく必要がある。

ソースコードをダウンロード後、展開したソースディレクトリで以下のようにビルドする。

```
cmake .
make
make install
```

- 使用方法

Moses:

1)対訳コーパスの原言語、目的言語の各テキストをトークナイザで単語もしくは形態素列に分割する。

2) clean-corpus-n.perl を用いて(1)で作成したデータをクリーニングする。

3) train-model.perl を用いて単語アラインメントを実施する。起動に際して各種パラメータの設定が必要であるが、詳細は利用環境の構築方法で述べたマニュアルに記載されている。

mgiza:

Moses で提供される訓練ツールである train-model.perl において、起動時のパラメータに -mgiza を加えることで使用する。併せて、並列処理のために使用する CPU の数を -mgiza-cpus NUMBER で指定する。

・ツールのコスト及びツールと依存関係になる他のソフトのコスト

ツール本体ならびに動作に必要な各種ソフトは無料で入手できる。

・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）

なし。

・ツールの特性（処理誤りの傾向等）

特記事項なし。

・ツールの処理品質

出力結果の品質は訓練データの量に依存するものであり、ツールの処理品質が問題となることは本調査においてはなかった。

・ツールに入力したデータ量と出力された辞書・コーパス等の量

表 2.1-8 を参照。

・処理に要する演算コスト（マシン性能と処理時間）

本調査で訓練データ数が最も多い越英の場合で、Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz の CPU を最大 8 並列で動作させ 10.6 分。

・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

言語依存性はないためベトナム・タイ語で利用可能。

２．２　タイ語用ツールの調査

２．２．１　OCR ツールの調査

（１）調査対象ツール

タイ語の OCR ツールとして調査を行ったのは以下の２つのツールである[23]。

(a)ABBYY 社 FineReader

(b)Nuance 社 OmniPage Server

ABBYY FineReader は PC 用のパッケージソフトである。本ソフトは、処理対象であるイメージデータの傾きの修正やテキスト化したい領域の指定等を対話的に行うための機能を備えている。一方 Nuance OmniPage Server は、PC やタブレット端末などのクライアントからサーバにイメージデータを送付し文字認識結果を受け取る形で利用するタイプのソフトウェアである[24]。個々のクライアントに OCR ソフトをインストールする必要がなく、また計算資源が大きくないハードウェアのクライアントからも OCR を気軽に利用できるというメリットはあるが、ABBYY FineReader のように、イメージの特定部分だけを読み取るといったことができない。もし行いたい場合には、利用者自身で特定の範囲だけを切り出したイメージデータを新たに作る作業が必要となる。

（２）評価

（２－１）評価で使用したデータ

特許庁から貸与されたタイ語公報全文イメージデータの中から日本語と英語でも出願されたパテントファミリーを持つもの計 20 件を候補として選定した。左記 20 件のうち、10 件は日本の法人が出願したもの、残り 10 件は米国の法人が出願したものである。また読み取り精度の評価にあたって、特定の公報だけ極端に文字数が多くなるのを避けるため、それぞれの案件において読み取るページ数は最大 13 ページとし、20 件全体での平均は 11.4 ページで、総ページ数は 228 ページとした。

（２－２）評価手順

評価の手順は以下の通りである。

①各 OCR ツールにてイメージデータを一括で読み取る。

②読み取りによって得られた OCR テキストにおいて、本来の読み取り対象であるテキスト以外に一緒に読み取られてしまった行番号等を削除する。（図 2.2-1 参照）

---

[23] 調査候補の選定にあたりタイの研究者から ArmThai という OCR ツールの推薦を受けたが、インターネット上にある同ソフトの紹介ページでソフトをダウンロードするリンクが無効となっていたため、実際にソフトを試用して評価することはできなかった。

[24] OmniPage にも PC 用のパッケージソフトが存在するが、当該ソフトにはタイ語の認識機能が含まれていなかったため、領域指定等の機能のないサーバ版を使用せざるを得なかった。

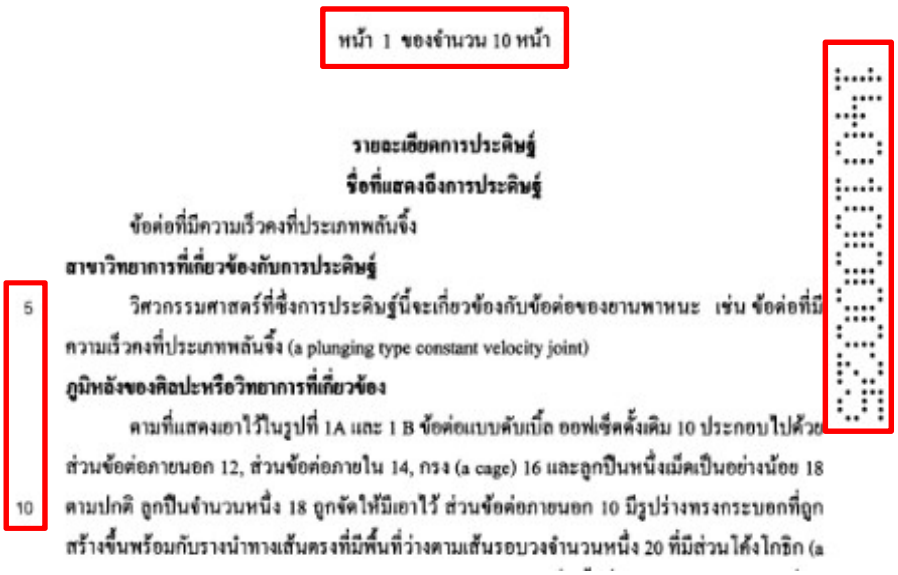


図 2.2-1 タイ語公報のイメージデータの一例

（赤枠の部分が左から、行番号、ページ番号、出願番号を示す。）

③OCR における文字認識の過程において、テキスト化すべき領域と判定されず全くテキスト化されなかった領域を手動で指定して再度読み取りを行い、既存の読み取り結果に追加する。ただし本作業は読み取り領域を対話的に指定できるインタフェースを有するツールでないと作業コストが非常に大きいため、本調査においては ABBYY FineReader でのみ実施する。

④ABBYY FineReader の OCR テキストにおいて誤認識された文字を修正し正解データを作成する。

⑤両 OCR テキストを正解データと比較し、正解率を算出する。

（２－３）評価結果

表 2.2-1 に二つのツールによる OCR テキストの正解率を示す。全データに対しての正解率は ABBYY FineReader が 96.65%、Nuance OmniPage Server が 92.37%であった[25]。

[25] 正解率の算出にあたっては、今回実施した文書全体での正解率を算出する以外に、OCR でテキスト化したデータの中で最終的な対訳コーパスに使用されたものだけを対象として算出することも可能である。文献#1-5 の 5 文献で算出したところ、対応する日本語文が見つからなかったタイ語文の全文字数は 2,694 文字でそのうち誤認識文字数は 123 文字であり、それらを除外して正解率を算出したところ 5 文献平均で 95.74%だったものが 95.75%となり、ほぼ同じであった。

表 2.2-1 文献別精度

#	出願番号	IPC	作業頁数	文字数（正解）	ABBYY FineReader		Nuance OmniPage	
					誤認識文字数	OCR正解率	誤認識文字数	OCR正解率
1	1401000025	F16D 3/226	10	18,282	597	96.73%	1,309	92.84%
2	1401000159	B68G 5/02	13	18,223	668	96.33%	1,612	91.15%
3	1401000186	H02H 3/20	13	22,115	774	96.50%	1,645	92.56%
4	1401000515	B23P 11/00	9	10,786	378	96.50%	578	94.64%
5	1401000595	C02F 1/20	13	22,637	1,501	93.37%	2,554	88.72%
6	1401000603	F24C 7/02	8	13,680	163	98.81%	1,709	87.51%
7	1401000625	B65D 71/20	11	17,803	761	95.73%	1,442	91.90%
8	1401000657	B60K 6/40	13	21,891	416	98.10%	1,404	93.59%
9	1401000713	C12N 1/02	13	19,850	830	95.82%	1,448	92.71%
10	1401000790	A61J 3/10	13	23,123	692	97.01%	968	95.81%
11	1401001130	F16H 48/08	7	10,769	171	98.41%	563	94.77%
12	1401001224	B01D 39/16	13	22,121	1,255	94.33%	1,599	92.77%
13	1401001236	G06F 15/16	6	9,581	212	97.79%	716	92.53%
14	1401001665	B65D 17/00	12	21,486	587	97.27%	1,507	92.99%
15	1401002175	F16D 13/52	13	18,529	522	97.18%	2,953	84.06%
16	1401002243	A47J 27/16	13	20,605	805	96.09%	1,488	92.78%
17	1401002340	H03M 7/40	10	13,499	319	97.64%	866	93.58%
18	1401002344	E04D 3/362	12	22,308	537	97.59%	1,390	93.77%
19	1401002566	C12P 7/50	13	22,690	644	97.16%	1,472	93.51%
20	1401003063	G01F 1/692	13	20,616	582	97.18%	1,067	94.82%
計			228	370,594	12,414	96.65%	28,290	92.37%

（3）考察

（3－1）OCR ツールの比較

（2－2）③で述べたように、最初の一括読み取りで画像領域として認識され、テキスト化されなかったテキストに関し、ABBYY FineReader ではその領域を対話的に文字認識させることができた。しかし Nuance OmniPage Server では機能上それを行うことができなかった。検索を目的とした場合であれば、テキスト化できない部分があっても読み取れた部分だけを検索用のインデックス作成に使えば問題とならない場合もあるが、対訳コーパス作成の場合は影響が大きい。すなわち英語や日本の明細書で対応する文が見つけられないため、読み取れなかった部分だけでなく、その前後の文対応を間違う可能性が高くなる。OCR を対訳コーパスの作成目的で使用する場合には、このようなテキスト化の抜けは可能な限りなくすことが望ましく、領域を指定した読み取り機能が必須であると思われる。

（３－２）OCR ツールの効率的な使い方

今回の調査では、計 228 ページの公報イメージデータを二つのツールで読み取る作業を行う必要があったため、テキストとして読み取る領域を指定できる OCR ツールにおいても全てのイメージデータを最初に一括で読み取り、そのあとで行番号などの本文以外の文字を削除する後処理作業を行った。しかし、数字として認識される行番号と違い、タイ語公報右上部にある出願番号は、印刷した明細書全ページに対して穴文字を用いて契印されたもので、印刷された字体とは大きく異なりしかも横向きである。そのため、領域を指定しないで読み取りを行うと、それらは全て行末尾にノイズとして付加されてしまい、削除するのに非常に手間がかかる。したがって、タイ語公報データを OCR で読み取るには、1 ページごとにテキスト化する領域を指定して対話的に認識処理を行うか、そうでなければ認識処理を行う前に画像編集ソフトを用いてテキスト化する領域のイメージデータをオリジナルデータから切り出す、もしくはノイズとなる穴文字の部分のイメージを消去してテキスト化したい領域だけのイメージデータを作ってから一括で認識処理を行った方が、テキスト化作業の効率が上がると考えられる。

（３－３）辞書登録による精度改善

ABBYY FineReader には辞書登録機能があるため、小規模ではあるが辞書登録による精度改善の効果を検証した。使用したデータは（２－３）で評価に用いたページの中の 1 ページで、認識誤りのあった文字を含む語を 21 語登録した。その結果、83 字の読み取り誤りが 27 文字減って 56 字となり、同ページの正解率は 94.4%から 96.2%へと改善した[26]。費用対効果を考えると全ての読み取り対象に対して同程度の辞書登録を行うことは現実的には難しいが、特許文書において出現頻度の高い用語を辞書登録することでテキスト化作業の効率を改善することが期待できる。

２．２．２　文分割ツールの調査

（１）調査対象ツール

タイ語の文分割ツールとして研究目的以外に利用可能と思われるツールは調査開始時点で OpenNLP しかなかったため、同ツールのみを調査した。

（１－１）Apache OpenNLP

・利用条件

Apache(v2)ライセンスの下で利用できる。

[26] 対象ページにおける 83 文字の読み取り誤りのうち、登録した 21 語に含まれる文字は 41 文字であった。そのうち辞書登録によって誤りが解消された文字は 32 文字である。ページ全体で削減された誤りは 27 文字であるので、5 文字は辞書登録後の読み取りで新たに発生した誤りとなる。

・入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

https://opennlp.apache.org/cgi-bin/download.cgi

・利用環境の構築方法

Java で実装されているため、利用するには Java のインストールが必要である。

・使用方法

以下のように Java の起動時にパラメータとしてタイ語の文分割機能と、使用するバイナリデータ（gz 圧縮形式）を指定する。

java opennlp.tools.lang.thai.SentenceDetector ${MODEL_DIR}/thai.sent.bin.gz

処理対象の原文の入力と処理結果の出力は標準入出力を通して行う。

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

Java が必要であるが、無料で入手できる。

（2）評価

タイ語では他の多くの言語と違い、文末を示すピリオドや句点の類が存在しない。そのため、文のどの個所でも分割される可能性があることになり、他の言語の文分割精度のような高い処理精度は期待できないと考えられる。実際２．１．２で述べたベトナム語用文分割ツールの精度と比べても低い精度となっている。

表 2.2-2 に文献別の評価結果を示す[27]。表中、「未分割」とは文末と判定して文分割を行うべき個所で分割しなかったケース、「誤分割」とは文の途中であるため文分割してはいけない個所で分割されたケースを示す。表から明らかなように、使用した文分割ツールの誤りは誤分割が多かった。今回評価に使用した文の総数は 3,018 文であるので、ツールの正解率は以下のようになる[28]。

1-(676+1,878)/3,018=15.4%

[27] 表中#17 については、タイ語公報のイメージデータにおいて、明細書に使用されるファイル名で請求項のイメージデータが格納されており、明細書のイメージデータが存在しなかったため欠番とし、明細書のデータ#21 と差し替えを行った。請求項については、パテントファミリーであっても出願される国毎に請求項の数や内容が異なる可能性が高く、文対応が困難となる可能性が高い。事実#17 では、タイ語と英語で請求項の数が異なっていた。

[28] 正解率の算出は本調査事業の仕様書に記載された算出式「1-（修正箇所／全文数)」に基づく。

表 2.2-2 文分割の結果（タイ語）

#	出願番号(タイ)	文数 (正解)	未分割	誤分割	修正計
1	1401000025	141	50	44	94
2	1401000159	248	12	264	276
3	1401000186	184	98	60	158
4	1401000515	81	14	39	53
5	1401000595	200	50	118	168
6	1401000603	128	41	51	92
7	1401000625	168	66	41	107
8	1401000657	143	37	120	157
9	1401000713	195	30	105	135
10	1401000790	168	37	72	109
11	1401001130	103	44	28	72
12	1401001224	182	14	70	84
13	1401001236	81	21	40	61
14	1401001665	151	45	81	126
15	1401002175	109	17	144	161
16	1401002243	143	15	81	96
17	欠番				
18	1401002344	126	37	152	189
19	1401002566	173	16	86	102
20	1401003063	199	18	190	208
21	1001001739	95	14	92	106
計		3,018	676	1,878	2,554

なお、単純な誤り数の比較では、誤分割が未分割の約 2.8 倍となっているが、文分割の誤りがあるとそれだけで直ちに最終的な目標である対訳コーパスや対訳辞書の精度に大きく影響するというわけではない。この後の文アライメントの処理において 1 文対 1 文(以下「1 対 1」のように「文」を省略して記す)の対応だけでなく、1 対多もしくは多対多の対応付けが行われる可能性があるためである。説明の便宜上、以下では英語と日本語のアライメントのために英語の文分割を例に説明する。

下の例文 2.2-1 において、仮に全てのピリオドを文末と誤って認定すると、例 2.2-2 のように全部で 7 つのフラグメントに分割されてしまう（”///”が文の切れ目を示す)。し

かし文アライメントの段階で、これらのフラグメント全体で「この手法はウイリアム博士により 2013 年 6 月発行の ACM J. Data Inf. Qual.で紹介された。」という日本語 1 文に対応すると推定されれば、英語の誤分割は意味を失う。

例 2.2-1　The method was introduced in ACM J. Data Inf. Qual., Vol. 10, No. 6, 2013 by Dr. Williams.

例 2.2-2　The method was introduced in ACM J. /// Data Inf. /// Qual. /// , Vol. /// 10, No. /// 6, 2013 by Dr. /// Williams.

一方、未分割についても、テキストの対応付けという意味では誤分割と同じである。例えば下の例の 2.2-3 では、本来セミコロン後で分割されるべき 2 文であるが、仮に分割されなかったとする。

例 2.2-3　FIG. 1 is an elevational view showing the suspended plate rack of the present invention; FIG. 2 is a view taken along line 2--2 of FIG. 1.

文アライメント処理において、この 1 文が「図 1 は、本発明の吊下板ラックを示す正面図である。図 2 は、図 1 の 2--2 線に沿った断面図である。」という日本語 2 文に対応すると推定されれば、文の対応付けとしては必ずしも間違っているとは言えないが、その後文アライメントされた対訳テキストを用いて Moses によりフレーズテーブルを作成する際に、本来の 1 文単位で処理する場合に比べ、2 文対 2 文の対訳での対応付けの方が個々の単語に対応する単語の候補数が増えるため、対応付けの精度が低くなる可能性がある。そのため、定性的には未分割の誤りよりも誤分割の誤りの方が辞書作成への悪影響は少ないと考えられる[29]。

（3）まとめ

評価した OpenNLP の処理精度以外の情報をまとめる。

（３－１）OpenNLP(タイ語文分割ツールとして)

・ツールの入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

https://opennlp.apache.org/cgi-bin/download.cgi

[29] 未分割が 50 か所、誤分割が 44 か所ある、修正前の文書(#1)を用いて文アライメントを実施したところ、分割箇所の誤りを修正した後のテキストを用いた場合のアライメント結果と比較して、未分割箇所の 23 か所、誤分割箇所の 17 か所はアライメント精度の直接的な悪化要因とはなっていなかった。したがって文書#1 での正解率は本来 1-(50+44)/141=0.33 であるが、実質的には 1-(50-23+44-17)/141=0.72 であるとみなすこともできる。

・ツールの利用環境の構築方法、使用方法

Java で実装されているため、利用するには Java のインストールが必要である。

使用する際は、以下のように Java の起動時にパラメータとしてタイ語の文分割機能と、使用するバイナリデータ（gz 圧縮形式）を指定する。

java opennlp.tools.lang.thai.SentenceDetector ${MODEL_DIR}/thai.sent.bin.gz

処理対象の原文の入力と処理結果の出力は標準入出力を通して行う。

・ツールのコスト及びツールと依存関係になる他のソフトのコスト

ツール本体ならびに動作に必要な Java はどちらも無料で入手できる。

・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）

使用していない。

・ツールの特性（処理誤りの傾向等）

本来分割すべきでない位置で誤って分割する傾向が強い。

・ツールの処理品質

句読点が使用されないタイ語の性質上仕方がないが誤りが非常に多く、処理品質は低い。

・処理に要する演算コスト（マシン性能と処理時間）

CPU が Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz×40、メモリが 396GB のマシンで 20 ファイル(計 3,018 文)の処理に 4.3 秒[30]。

・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

タイ語用のツールであるため対応可能。

２．２．３　トークナイザの調査

（１）調査対象ツール

タイ語のトークナイザとしては、文分割ツールとしても評価した OpenNLP と、SWATH と呼ばれるツールの二つを調査した。

（１－１）Apache OpenNLP

・利用条件

Apache(v2)ライセンスの下で利用できる。

・入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

[30] 3 回の実行時間の平均。初回は 5.7 秒で 2 回目以降はキャッシュにより 3.7 秒、3.6 秒と大きく異なる値となったが、ここでは単純に平均を記載する。また 1 ファイルごとにコマンドを起動し 20 ファイル処理する時間であるため、3,018 文を 1 ファイルにまとめて処理する場合には処理時間は変わってくる。以下同様。

https://opennlp.apache.org/cgi-bin/download.cgi

・利用環境の構築方法

Java で実装されているため、利用するには Java のインストールが必要である。

・使用方法

以下のように Java の起動時にパラメータとしてタイ語のトークナイズ機能と、使用するバイナリデータ（gz 圧縮形式）を指定する。

java opennlp.tools.lang.thai.Tokenizer ${MODEL_DIR}/thai.tok.bin.gz

処理対象の原文の入力と処理結果の出力は標準入出力を通して行う。

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

Java が必要であるが、無料で入手できる。

（１－２）SWATH

・利用条件

GNU General Public License の下で利用できる。

・入手方法

以下の URL にてソースコードをダウンロードする(2017 年 3 月末現在)。

ftp://linux.thai.net/pub/thailinux/cvs/software/swath

・利用環境の構築方法

ダウンロード後展開した同ツールのディレクトリにて、configure ならびに make を実行する。

・使用方法

処理対象の原文の入力と処理結果の出力は標準入出力を通して行う。

原文データのフォーマットとして通常のテキストの他、html や rtf などがサポートされており、それらをパラメータで指定することができる。また単語分割のアルゴリズムとして辞書との最長一致や bigram に基づく計４種類の方式が指定できる。

・ツールのコスト

無料

・ツールと依存関係になる他のソフトのコスト

他ツールとの依存関係はないため追加コストはない。

（２）評価

トークナイザはコーパス作成やフレーズテーブル作成の過程で用いるツールである。このため、ここでの評価は、これらの過程で OpenNLP と SWATH のどちらを用いるべきかを判断するため、語分割の傾向を把握するための簡単な比較を行うに止めた。すなわち、OpenNLP

と SWATH によるトークナイズ結果を比較すると明らかに OpenNLP の方が短単位に分割しており、分割数が多かった。分割単位が長いとその語に対する出現頻度が短い場合と比べて小さくなるため、本調査のように対訳コーパスのサイズが小さい状況においては OpenNLP が出力する短単位の結果の方が適していると考え、以下の評価作業では同ツールを用いることとした。なお特許庁の貸与データから抽出した 30 万文のタイ語の特許要約文を用いて動作確認を実施したが、特に問題なく処理を行うことができた。

（３）まとめ

評価した OpenNLP の処理精度以外の情報をまとめる。

（３－１）OpenNLP(タイ語トークナイザとして)

・ツールの入手方法

以下の URL にてダウロード可能(2017 年 3 月末現在)。

https://opennlp.apache.org/cgi-bin/download.cgi

・ツールの利用環境の構築方法、使用方法

Java で実装されているため、利用するには Java のインストールが必要である。

使用する際は、以下のように Java の起動時にパラメータとしてタイ語のトークナイズ機能と、使用するバイナリデータ（gz 圧縮形式）を指定する。

java opennlp.tools.lang.thai.Tokenizer　${MODEL_DIR}/thai.tok.bin.gz

処理対象の原文の入力と処理結果の出力は標準入出力を通して行う。

・ツールのコスト及びツールと依存関係になる他のソフトのコスト

ツール本体ならびに動作に必要な Java はどちらも無料で入手できる。

・ツールを利用するに際して使用した学習用のデータ（データの種類・分量）

使用していない。

・ツールの特性（処理誤りの傾向等）

特になし。

・ツールの処理品質

ツール単体でのトークナイズの精度は評価できていない。

・処理に要する演算コスト（マシン性能と処理時間）

CPU が Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz×40、メモリが 396GB のマシンで 2,669 文の処理に 1.9 秒。

・カスタマイズ性（ベトナム・タイ語に対応するためのカスタマイズが容易か等）

タイ語用のツールであるため対応可能。

２．２．４　コーパス作成ツールの調査

（１）利用可能なツールの調査

調査対象ツールは、ベトナム語の場合と同じく、内山らのツールのみである。調査結果については、２．１．４を参照されたい。

（２）評価

２．１．４で示した越日、越英対訳コーパスを作成したのと同様の手順で作成した泰日、泰英対訳コーパスからそれぞれ 200 文をランダムに抽出し、文アライメント精度を評価した。評価結果は以下の通りである。ベトナム語の場合には英語とのアライメントの方が精度が高かったが、タイ語については日本語との対応の方が 6%ほど高い結果となっている。

表 2.2-3 文アライメント結果の評価（泰日、泰英）

スコア	泰日	泰英
A	172（86.0%）	159（79.5%）
B	12（6.0%）	16（8.0%）
C	16（8.0%）	25（12.5%）

また人手によるチェックを経て、最終的に得られた対訳コーパスのサイズは以下の通りである。

泰日：　2,669

泰英：　2,853

（３）考察

（３－１）アライメントツールのスコアと評価のスコアについて

２．１．４（３－２）での考察と同様に、タイ語の文アライメント結果についても、200 事例の評価結果を用いて、アライメントツールのスコアと評価のスコアの関係を分析し、大まかな傾向を把握することにする。

下の図は、文アライメントツールのスコアに対して、どのような品質評価基準のスコアが付与されたかを示したものである。それぞれの言語対での文アライメント結果をアライメントツールのスコアが 0.4 以上、0.3 以上 0.4 未満、0.2 以上 0.3 未満、0.1 以上 0.2 未満、0.1 未満の 5 つにグループ分けし、人手による評価のスコア A,B,C がそれぞれ何文対あるかを示している。ベトナム語の場合と異なり、(a)泰日や(b)泰英においてはスコア 0.1 未満の場合にも文対応の評価で A 評価となっている正しい文対応が多数含まれていることが分かる。

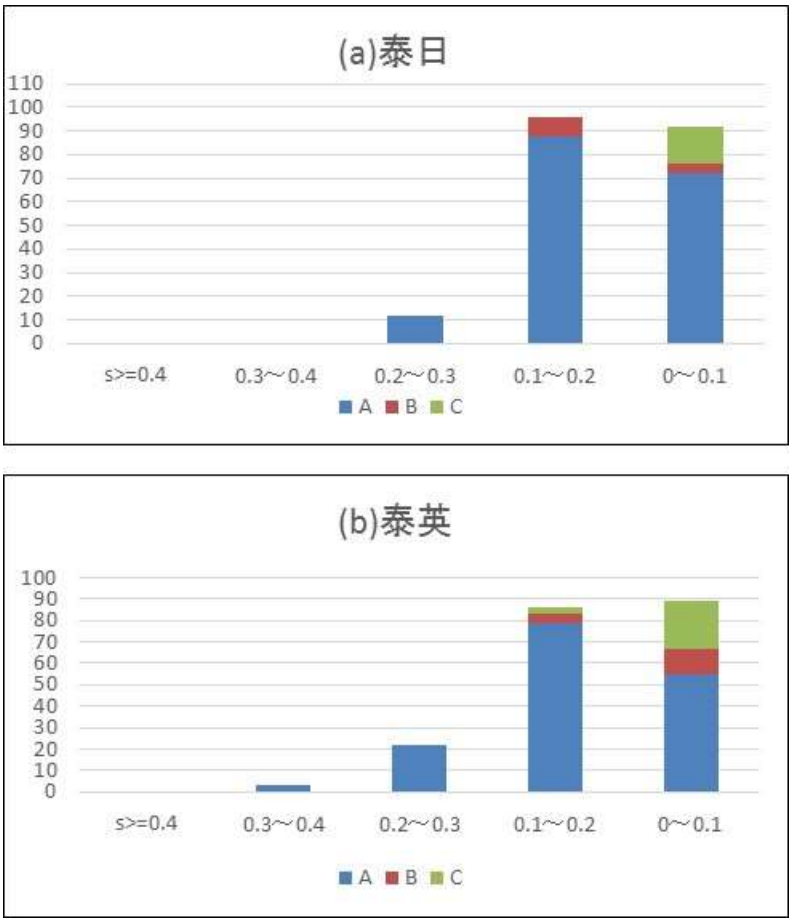


図 2.2-2 文アライメントスコアに対する文対応の評価結果

そこで、ベトナム語での場合と同様に、スコア A の割合が文アライメントのスコアに応じてどのように変化するかを調べた。図 2.2-3 がその結果である。この図より(a)泰日では文アライメントのスコアが 0.06 以上であればスコア A の割合がほぼ 90%以上となる。一方(b)泰英では、スコア A の割合を 90%文程度とするにはアライメントのスコアは 0.10 以上とする必要があり、仮にこの条件で文アライメント結果を絞り込むと、スコア A となる事例の約 35%(表 2.2-3 における泰英のスコア A の事例 159 件中 55 件)を取りこぼしてしまうことになる。

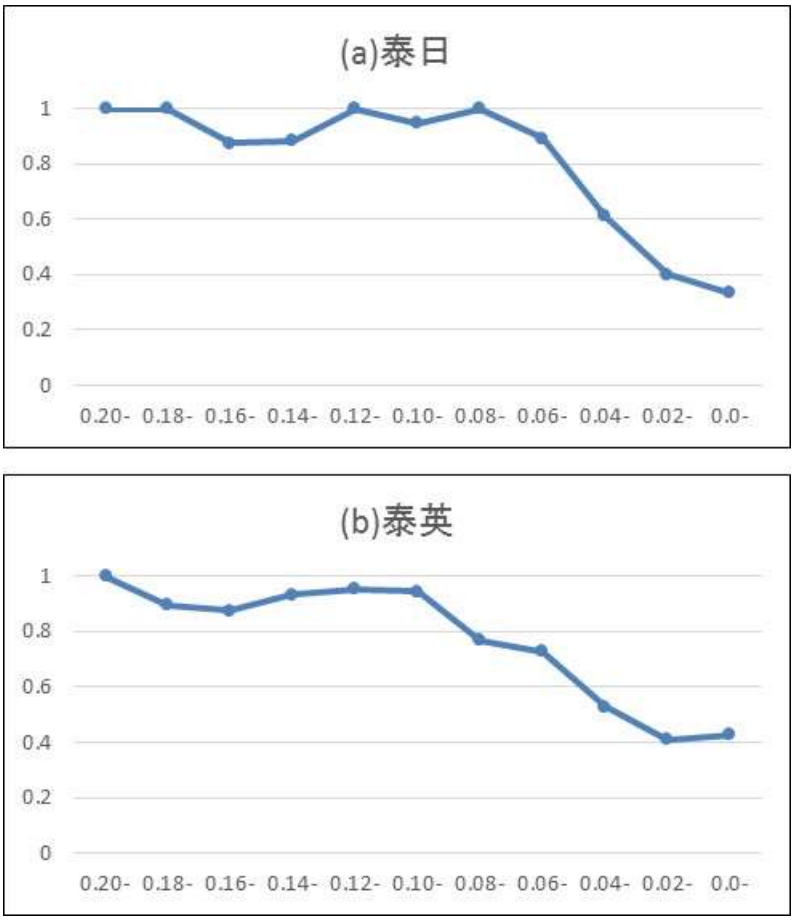


図 2.2-3 文アライメントスコアに対する文対応の評価結果(2)

続いて下表は、文アライメントにおいて、何文と何文が 1 つの対訳事例として抽出されたのか、そしてそれらがどのように評価されたかをまとめたものである。(a)泰日では、タイ語 1 文と日本語 1 文からなる文対が評価した 200 件のうち 94.0%の 188 件あり、それらのうち 165 件が A 評価であったことを示している。また 1:1 の対応事例においては A 評価が泰日 88%、泰英 84%と高い正解率が得られているが、1:1 以外の事例では A 評価が泰日 58%、泰英 25%と正解率が低く、1 対 1 対応以外の対訳を用いるには人手によるチェックが不可欠であると言える。

表 2.2-4 対応づけされた文数別のアライメント評価結果

(a)泰日

	A	B	C	計
1:1	165	11	12	188
1:2	2	0	0	2
2:1	4	1	1	6
4:1	1	0	0	1
5:1	0	0	3	3

(b)泰英

	A	B	C	計
1:1	155	14	15	184
2:1	4	1	4	9
3:1	0	0	1	1
5:1	0	1	5	6

（３－２）原出願の違いによる文アライメントの傾向

２．１．４（３－３）でのベトナム語での分析と同様、原出願が日本か米国かでアライメントの精度にどのような違いがあるかを分析した。図 2.2-4 がその結果である。ベトナム語の場合は、越日において原出願が日本の場合のアライメント精度が原出願が米国のものよりも低かったが、タイ語の場合は、泰日、泰英ともに原出願が日本の場合の文アライメント精度が低く、両者の間では泰英の方がより精度差が大きいことが分かった。

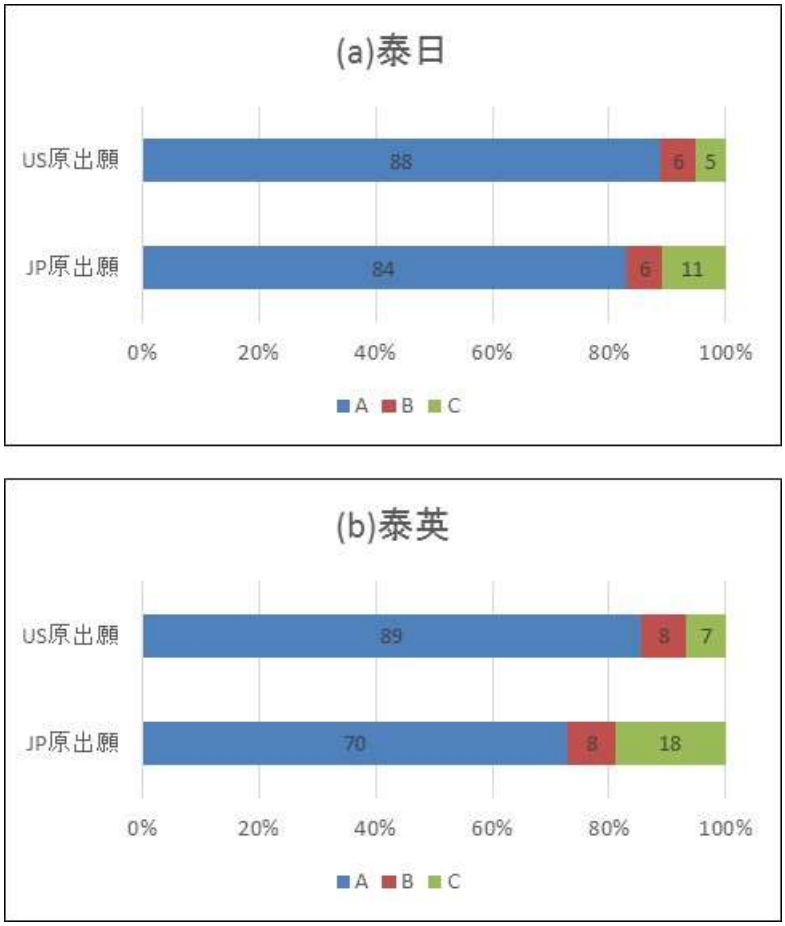


図 2.2-4 原出願国別の文アライメントの評価（泰↔日／英）

（４）まとめ

ツールの入手方法や使用方法、コスト等については２．１．４に記載の通りであり、ここではタイ語の文アライメントのみに関連する内容を述べる。

・ツールに入力したデータ量と出力された辞書・コーパス等の量

ツールに入力したデータと出力されたアライメントの数は以下の通りである[31]。

[31] タイ語の文数は、文分割ツールの処理結果を人手で修正した後の文の数であるが、日本

表 2.2-5 ツールに入力したデータ量と出力されたアライメント数

項目		泰日	泰英
入力	タイ語文数	3,018	3,018
	日本語文数	3,064	－
	英語文数	－	3,027
出力		2,617	2,555

・処理に要する演算コスト（マシン性能と処理時間）

泰日アライメント(20 文献)の処理を行うケースで、RedHat 上のバーチャルマシン(ホストマシンのスペック： CPU が Intel(R) Xeon(R) E5-2690 v2 @ 3.00GHz×20、メモリが 32GB)にて 6.9 分。

２．２．５　辞書作成ツールの調査

（１）利用可能なツールの調査

調査対象ツールは、ベトナム語の場合と同じく、Moses と mgiza である。調査結果については、２．１．５を参照されたい。

（２）評価

（２－１）辞書登録候補語の作成手順

作成手順は２．１．５に示したベトナム語の場合と同様である。

（２－２）パテントファミリーを用いた発明の名称の対訳コーパスの作成

今回の評価で対訳コーパスの試作に用いる明細書は、泰英・対日対訳の場合 20 文書しかなく、対訳コーパスのサイズで見ても泰日が 2,669、泰英が 2,853 とそれほど多くない。そこでベトナム語の場合と同様、特許ドメインでの対訳コーパス量を増やすため、パテントファミリーの発明の名称を抽出して対訳コーパスを作成した。

タイ語の特許に対応する日本語ならびに英語のパテントファミリーの発明の名称は、特許庁から貸与されたタイ語公報のデータを利用して以下のように抽出した。

①貸与されたデータの中に書誌事項が記載された以下の例のような XML ファイルが含まれているので、それからタイ語のタイトルと優先権主張番号に JP（または US）を含むもののペアを取り出す。

語と英語の文の数は本ツールの文分割機能によって分割された文の数であり、誤りを含んでいる可能性がある。そのため、実際の文の数はこれらとは異なる可能性がある。

```
<?xml version="1.0"?>
<th-patent-document　・・・>
  <bibliographic-data id="bibl" lang="TH" country="TH">
    (中略)
    <priority-claims>
      <priority-claim kind="national">
        <doc-number>PCT/JP2012/004388</doc-number>
        <date>2012-07-05</date>
      </priority-claim>
    </priority-claims>
    <invention-title lang="TH">
ชุดเครื่องป้อนอนุภาคและชุดเครื่องผลิตผลิตภัณฑ์แผ่น</invention-title>
  </bibliographic-data>
  <abstract id="abstract" lang="TH">
(以下省略)
```

図 2.2-5 タイ語公報データの一例

②優先権主張番号をもとに日本語明細書を検索し日本語のタイトルを取得する。(英語も同様)

以上の手順により、以下の件数のパテントファミリーが得られ、各言語での発明の名称を用いて同件数の対訳コーパスが得られた。

泰日： 15,128 件

泰英： 9,726 件

（２－３）日泰、英泰辞書の作成

２．１．５で述べた日越、英越辞書と同様に、日泰、英泰辞書を作成した。

作成の過程で得られたデータ数、ならびに最終的に得られた辞書登録語候補の数は以下の通りである。

表 2.2-6 日泰、英泰辞書作成に関連するデータ数

		日泰	英泰
訓練データ数(クリーニング前)		95,376	147,202
内訳	明細書対訳	2,669	2,853
	名称対訳	15,128	9,726
	対訳辞書[32]	77,579	134,623
訓練データ数(クリーニング後)		95,031	146,976
フレーズ数		573,303	669,470
言選が抽出した用語数		3,999	5,518
フレーズテーブルから抽出できた辞書登録候補語の数		2,468	2,641

得られた日泰、英泰辞書を日越、英越辞書の場合と同様に、本調査で定められた品質評価基準に基づき先頭 200 語を評価した。その結果を図 2.2-6 及び図 2.2-7 に示す。日泰辞書の第 1 訳語はスコア A が 82.5%、英泰辞書は 74.0%となり[33]、日越・英越辞書の結果でも述べたように、この精度で大規模な抽出ができれば実用に供するレベルであると判断できる。

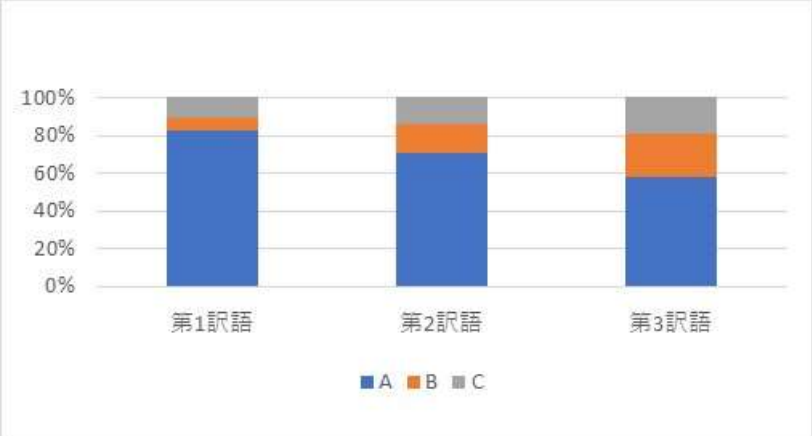


図 2.2-6 日泰辞書精度

[32] ここで用いた対訳辞書は３．４．２で作成した Wikipedia のデータを用いて試作した対訳辞書である。

[33] ２．２．４で作成した泰日、泰英対訳コーパスにおいて共通のタイ語文を持つ事例から原出願が日本と北米の事例各 50 文を選択し、その中のタイ語一語に対応する語が英語文ならびに日本語文に存在するかを調べたところ、日本が原出願の場合は泰日、泰英合わせて 95 語の対応が取れたのに対し、北米が原出願の場合は泰日、泰英合わせて 84 語しか対応が取れず、原出願が日本の場合の方が単語の対応が取りやすい傾向があった。ただし、泰日と泰英では泰英の方が単語の対応度は高く、辞書の精度評価とは異なる傾向を示した。

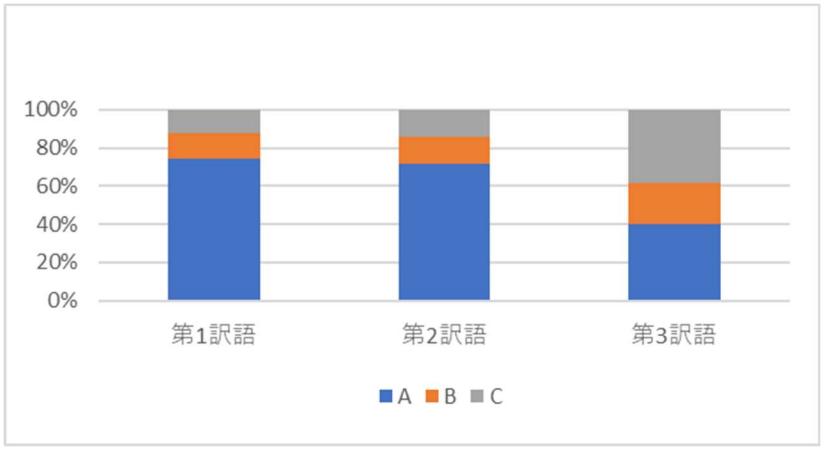


図 2.2-7 英泰辞書精度

３．パテントファミリー以外から対訳コーパス・辞書を作成する方法の検討

本章では、法的・技術的な観点から、パテントファミリー以外の言語資源を用いて対訳コーパスや辞書を作成する手法を検討し、可能な手法についてはさらに次の観点で調査を行う。

・作成可能な対訳コーパス・辞書の分量
・作成される対訳コーパス・辞書の特徴
・作成の具体的手順（作成の準備含む）
・作成に要する期間（作成の準備含む）
・作成コスト
・作成に際して存在する課題

なお、以下の記述にあたっては、説明の便宜上、ベトナム語とタイ語で共通する内容に関しては、ベトナム語を代表として使用する。タイ語に関しては、ベトナム語の場合とは異なる特筆すべき内容がある場合にのみ説明を追加することとする。また、対訳コーパスや対訳辞書を作成する場合でも、それらは機械翻訳システムの学習目的でのみ使うものとし、作成したコーパスや辞書を第三者に配布提供することはないという利用形態を制約条件として置くこととする。

３．１　英語を中間言語とした辞書作成

英語を中間言語とした辞書作成とは、既存のベトナム語と英語との対訳辞書と、英日辞書を総合することで越日対訳辞書を構築する方法である。

（1）法的・技術的な観点から見た方法の妥当性

法的には、使用する越英辞書ならびに英日辞書において使用にあたっての著作権上の制約条件がなければ、それらを組み合わせて新たな越日辞書を作って使用することは問題ないと考えられる。

技術的には、本構築手法は越英辞書の訳語である英語の見出しが英日辞書にあるかを確認し、あればその日本語訳を元のベトナム語の訳語とすることでベトナム語と日本語の対応関係を求めることで作ることができる。しかしながら、この対応関係は２つの辞書の単純な組み合わせであるので、最終的に得られるベトナム語と日本語のペアが、対訳としては妥当性が低いような組み合わせが生成される可能性もある。訳が一意に決まるような専門用語の場合には問題になる可能性は低いが、一般的に複数の語義を持つ基本的な用語を含む辞書を作る場合には、可能な訳を単に抽出するだけでなく、最も標準的な訳語を決めることができれば、より有益な辞書とすることができる。また、利用する辞書に品詞情報があれば、英訳語が多品詞であっても、その品詞に合致した適切な日本語訳を割り当てられる可能性がある。

（２）調査の観点での検討

（２－１）作成可能な辞書の分量

英日辞書には大規模な汎用の辞書が各種存在するため、越英辞書の訳語として記載されている英語は、英日辞書に収録されている可能性が高い。そのため、本手法で作成可能な辞書の分量は、基本的に越英辞書の語彙数によって決まると考えられる。

（２－２）作成される辞書の特徴

作成される辞書の質は、利用する越英・英日辞書の質に依存する。従来のように人手で編集して作られた辞書であれば十分高い質が期待できるが、最近では統計処理により作っている辞書も存在するため、辞書を組み合わせて作る本手法の場合には、それぞれの辞書の質をきちんと把握することが肝要と思われる。

（２－３）作成の具体的手順（作成の準備含む）

英日辞書を使う場合の手順は以下の通りである。

a) 越英、英日辞書を入手する。
b) 英日辞書を指定した英語見出しで検索できるようにする。
c) 越英辞書からベトナム語見出しと英語訳語を取り出し、取り出した英語に対してｂ）で準備した英日辞書検索機能を用いて日本語訳語を取り出し、越日辞書項目として出力する。

英日辞書の代わりに英日機械翻訳システムを使う場合の手順は以下の通りである。

a) 越英辞書と英日機械翻訳システムを入手する。
b) 越英辞書からベトナム語見出しと英語訳語を取り出し、取り出した英語を英日機械翻訳システムで日本語に翻訳し、越日辞書項目として出力する。
c) 英日機械翻訳システムによる翻訳結果に原文の英語が未知語として残っている場合には、当該英語は翻訳システムの英日辞書に登録されていないので、ｂ）で出力された越日辞書項目から削除する。

（２－４）作成に要する期間（作成の準備含む）

単純な辞書の組み合わせでベトナム語と日本語のペアを出力するだけであれば、作成時間はほとんどかからない。それぞれの辞書がテキストデータとして与えられる場合、データのフォーマットを確認し、辞書データをパーズするプログラムが必要であるが、データ形式が XML や JSON 等であれば、それも既存のツールやライブラリを利用して短時間で実現可能である。

（２－５）作成コスト

越英辞書ならびに英日辞書に有料のものを導入する場合には、そのコストがかかる。また英日辞書データの代わりに英日機械翻訳ソフトを導入する場合は、そのコストがかかる。それ以外の作成用ソフトは小規模であるので、そのコストは小さい。

（２－６）作成に際して存在する課題

（１）に記載した事項を除き、一般的には特になし。

（３）具体例

実際の言語資源を用いて上で述べた手順で対訳辞書を作成した結果を示す。検討に使用した辞書は泰英辞書の LEXiTRON である。当該辞書はオープンソースとして提供されており任意の目的での利用が可能である。辞書データは XML に準ずるタグを用いた形式となっている。具体例を図 3.1-1 に示す。

<Doc>
<tsearch>กงสี</tsearch>
<tentry>กงสี</tentry>
<eentry>firm</eentry>
<tcat>N</tcat>
<tenglish>company; family; gathering</tenglish>
<tsyn>บริษัท, หุ้นส่วน, กองกลาง</tsyn>
<tsample>กฎของที่นี่คือจะขายเอาเงินเข้ากระเป๋าตัวเองไม่ได้ต้องเก็บไปรวมไว้ในกงสี</tsample>
<tdef>ของกลาง, กองกลางที่ใช้รวมกันสำหรับคนหมู่หนึ่งๆ (จ.ว่าบริษัททำการค้า, กิจการที่จัดเป็นสาธารณะ)</tdef>
<id>29331</id>
</Doc>

図 3.1-1 LEXiTRON 泰英辞書データの例

・作成可能な対訳辞書の分量

当辞書には 33,057 語のタイ語見出しが含まれており、タイ語と英語のペアは 38,839 ある。すなわち、単純には一見出しあたり 1.2 個の英訳語が付与されているということになる。この 38,839 の項目の英語を英日機械翻訳ソフトで翻訳し、未知語として日本語訳を付与できなかったものを除くと、最終的に対訳辞書の総エントリ数は 38,578 となった。

กงสี ||| 会社 ||| firm
กงสุล ||| 領事 ||| consul
กงสุลใหญ่ ||| 総領事 ||| consul general
กงเกวียน ||| カートの車輪 ||| wheels of a cart
กงเต๊ก ||| 中国の葬式 ||| Chinese funeral ceremony
กงเต๊ก ||| 煉獄から魂を解放する式 ||| ceremony of releasing soul from purgatory
กงไฉ่ ||| 漬けてあるキャベツ ||| pickled cabbage
กฎ ||| 規則 ||| regulation
กฎ ||| 支配 ||| rule

図 3.1-2 作成された対訳辞書の例
（泰日英の 3 言語辞書となっている）

・作成される対訳辞書の特徴
英日機械翻訳ソフトを用いて英語を日本語に翻訳したため、動詞などの用言の訳が終止形になっていない場合がある。
・作成の具体的手順
（２－３）で述べた手順のうち、英日機械翻訳ソフトを用いる場合の手順で実施した。
・作成に要する期間
半日程度。
・作成コスト
英日機械翻訳ソフトは所有しているものを使用したため、本作業のための追加作成コストは発生していない。
・作成に際して存在する課題
特に大きな問題ではないが、辞書データにおいてタイ語見出しと英語見出しの数が合わなかったため、両方がペアで取り出せるデータを使用した。

３．２　英語を中間言語とした対訳コーパス作成

英語を中間言語とした対訳コーパス作成とは、本調査においては既存のベトナム語と英語の対訳コーパスにおける英文を、英日機械翻訳で日本語に翻訳することで越日対訳コーパスを構築する方法である。

（1）法的・技術的な観点から見た方法の妥当性

本手法で作成された越日対訳コーパスが統計翻訳の学習目的にのみ用いられ、コーパス自体を配布することがなければ著作権侵害となる恐れはないので、法的な問題はないと考

えられる。

技術的には、越英対訳コーパスの英文を機械翻訳するだけで作ることができるため、実現性に関する問題はない。

（２）調査の観点での検討

（２－１）作成可能な対訳コーパスの分量

越英コーパスの英文を英日翻訳して越日コーパスを作るという性質上、本手法で作成可能な対訳コーパスの分量は、越英コーパスの分量が上限となる。さらに、英日機械翻訳の結果を、例えば未知語の有無や構文解析に成功したか、訳文の日本語を言語モデルで評価した結果などにより最終的に翻訳結果として採用するかを判断するような仕組みを導入すれば、翻訳結果を無条件に採用する場合よりも対訳コーパスの質は向上するが、分量は越英コーパスのサイズよりも小さくなる。

（２－２）作成される対訳コーパスの特徴

本手法により作成される対訳コーパスの一番の特徴は、越日コーパスの日本語データが機械翻訳の結果であるため、このコーパスを用いて学習される統計翻訳システムのフレーズテーブルには必然的に機械翻訳の誤りが含まれるということである[34]。本手法で開発した越日コーパスに仮にそのような誤りが含まれていたとしても、パテントファミリーを用いて作成される越日対訳コーパスの分量が少ない時点では、そのような誤りよりも正しいフレーズが学習できるというメリットが勝ると思われる。しかしパテントファミリーを用いて作成される越日対訳コーパスの分量が徐々に増えていくと、やがて誤りの問題が顕在化する。本手法で開発した越日コーパスを使用しつつ継続的にパテントファミリーを用いた越日コーパスを開発していく場合には、注意が必要である。

（２－３）作成の具体的手順（作成の準備含む）

a) 越英対訳コーパスと英日機械翻訳システムを入手する。
b) 越英対訳コーパスの英文を英日機械翻訳システムで翻訳し、英文と翻訳結果を入れ替える。

（２－４）作成に要する期間（作成の準備含む）

作成時間は英日機械翻訳システムによる翻訳時間にほぼ等しい。英日機械翻訳システムによる翻訳時間は、使用する機械翻訳システムの方式がルールベース翻訳か統計翻訳等か

[34] 機械翻訳結果を人手で後編集し機械翻訳の誤りをなくすことで、機械翻訳結果をそのまま使うよりも高品質な対訳コーパスを作成することができる。その場合、後述する「人手翻訳による対訳コーパス」の方法と比べコストと作成期間を削減できる可能性は高いが、機械翻訳結果をそのまま使う場合に比べると圧倒的に作成コストは大きくなり、作成に要する時間は長くなる。

により大きく変わる。又、使用する機械翻訳システムを、利用可能な計算機環境で同時にいくつ実行できるかによっても翻訳時間は大きく変化する。

（２－５）作成コスト

越英対訳コーパスならびに英日機械翻訳に有料のものを導入する場合には、そのコストがかかる。また、（２－４）でも述べたように、使用する機械翻訳システムの種類や作成にかける時間によって準備すべき計算機環境が異なるため、既存の計算機環境で対応できない場合にはその調達コストが必要となる。

（２－６）作成に際して存在する課題

英日機械翻訳システムによっては、翻訳する英文が長いと訳文が全く出力されなかったり、複数の部分に分割されて翻訳されるため、原文の行数と訳文の行数が異なってしまうことがある。そのような場合には、英文を翻訳結果で置換する際に、英文と翻訳結果が適切に対応していることを確認する必要がある。その方法としては、例えば、英日翻訳を行っても変化しない数字や記号などを英文一文ごとにセパレータとして加え、仮に英文が分割されて翻訳された場合でも確実に対応が取れるようにすればよい。それ以外の方法としては、英日翻訳ソフトを表計算ソフトにアドインできる場合には、英文を一文ごと表計算ソフトのセルに入れて翻訳することで、文の対応を保った形での翻訳が可能である。

（３）具体例

実際の言語資源を用いて上で述べた手順で越日対訳コーパスを作成した結果を示す。使用した越日対訳コーパスは EVBNews コーパスである。本コーパスはベトナムの情報技術大学がウイーン大学等と共同で開発した 80 万文対以上を含む英越対訳コーパスである EVBCorpus の一部で約 45,000 文対からなる。図 3.2-1 に利用した EVBCorpus の一部を、また図 3.2-2 に作成した対訳コーパスの一部を示す。

<corpus url='http://code.google.com/p/evbcorpus/'>EVBCorpus</corpus>
<author email='hungnq@uit.edu.vn'>Quoc-Hung Ngo, Werner Winiwarter</author>
<citation>Quoc-Hung Ngo, Werner Winiwarter, (2012). ″Building an English-Vietnamese Bilin
gual Corpus for Machine Translation″, International Conference on Asian Language Processi
ng 2012 (IALP 2012), pp. 157-160, Ha Noi, Vietnam</citation>
</head>
<text>
<spair id='1'>
<s id='en1'>What is a Fenqing ?</s>
<s id='vn1'>Fenqing là gì ?</s>
</spair>
<spair id='2'>
<s id='en2'>Fenqing is a Chinese word which literally means ″ angry youth ″ .</s>
<s id='vn2'>Fenqing là một từ tiếng Hoa mà nghĩa đen là ″ thanh niên phẫn nộ ″ .</s>
</spair>

図 3.2-1 EVBCorpus 越英対訳コーパスデータの例

Fenqing là gì ? ||| Fenqing は何ですか。 ||| What is a Fenqing ?
Fenqing là một từ tiếng Hoa mà nghĩa đen là ″ thanh niên phẫn nộ ″ . ||| Fenqing は中国の単語です、どれ、文字通りに、手段″腹を立てた若さ。」 ||| Fenqing is a Chinese word which literally means ″ angry youth ″ .
Từ này có nhiều cách dịch sang tiếng Anh như là thanh niên hoài nghi , thanh niên theo chủ nghĩa dân tộc , thanh niên cuồng loạn và thanh niên tức giận . ||| この単語は、皮肉な若さ、若い国家主義者、ヒステリーの若さおよび怒れる若者たちのような英語の中に多くの翻訳を持っています。 ||| This word has many translations in English such as cynical youth , young nationalists , hysterical youth and angry young men .
Cá nhân tôi thích gọi chúng là bọn thanh niên du thủ du thực hoặc thanh niên phẫn nộ và ngu dốt . ||| 私は、それらを暴徒若さあるいは無知な腹を立てた若さと呼ぶことが個人的に好きです。 ||| I personally like to call them mob youth or ignorant angry youth .

図 3.2-2 作成された対訳コーパスの例
（越日英の 3 言語対訳コーパスとなっている）

・作成可能な対訳コーパスの分量

もとの越英コーパスと同じ約 45,000 文対。

・作成される対訳コーパスの特徴

コーパスの名称から、本コーパスの元データはニュース記事と思われる。そのため、ベトナム語のテキストは文体が整った良質のデータと考えられるが、特許とはドメインが異なるため、このコーパスを特許翻訳向けに学習する対訳データとして用いたときの有効性は不明である。またルールベース方式の英日機械翻訳で英文を翻訳しているため、翻訳結果である日本語文が非文となっている場合がある。

・作成の具体的手順

（２－３）で述べた通り。使用した英日機械翻訳ソフトの自動文分割機能により原文の行数と訳文の行数が異なることがあるため、英文一文ごとに文の切れ目を示す記号を入れ、翻訳結果ではその記号を基に、分割されて翻訳された結果を 1 行にまとめる処理を行った。

・作成に要する期間

1 時間程度。

・作成コスト

英日機械翻訳ソフトは所有しているものを使用したため、本作業のための追加作成コストは発生していない。

・作成に際して存在する課題

使用した英日機械翻訳ソフトが PC 用のソフトであったため、英文の文字コードを SJIS に変換する必要があった。越日対訳コーパスは UTF8 で作成されていたため、文字コードを SJIS に変換する過程で対応する文字コードが SJIS にない文字は削除されている可能性がある。

３．３　TED からの対訳コーパス・辞書作成

TED からの対訳コーパス・辞書作成とは、TED Talks のテキストデータから対訳コーパスを作成し、該対訳コーパスから辞書を生成する方法である。

（1）法的・技術的な観点から見た方法の妥当性

本手法で作成された越日対訳コーパスが統計翻訳の学習目的にのみ用いられ、コーパス自体を配布することがなければ著作権侵害となる恐れはないので、法的な問題はないと考えられる。また当該対訳コーパスを用いて作成される対訳辞書も、元のコンテンツに対して統計的な解析を行って得られるものであって同コンテンツの著作物性を侵害する内容とはならないので、法的な問題はないと考えられる。

技術的には、日本語のデータで文末を示す句点が記述されているものが約 5%しかないと

いうことが文アライメントを行う上での課題である。約 95%のデータでは句点が使われていないため、そのままでは日本語の文末の正確な判定ができない。そこで映像とテキストの同期用の情報を利用する。TED の書き起こしデータとその翻訳テキストは講演の映像に対して字幕で表示する単位で区切られている。そこで、最初に英語側で文を認定し、それに対応する日本語およびベトナム語の翻訳テキストを取り出せば、それで文アライメントを行うことができる。(図 3.3-1 から 3 参照)

<seekvideo id="1157">In the past few months, I've been traveling for weeks at a time</seekvideo>
<seekvideo id="4532">with only one suitcase of clothes.</seekvideo>
<seekvideo id="6892">One day, I was invited to an important event,</seekvideo>
<seekvideo id="9075">and I wanted to wear something special and new for it.</seekvideo>
<seekvideo id="12140">So I looked through my suitcase and I couldn't find anything to wear.</seekvideo>

図 3.3-1　英語のデータの一例

<seekvideo id="1157">近頃よく 時には 1 度に何週間も スーツケース 1 つで旅に出ていました</seekvideo>
<seekvideo id="6892">あるとき 重要なイベントに 招待され</seekvideo>
<seekvideo id="9075">そのために何か特別な新しい服を 着たいと思い</seekvideo>
<seekvideo id="12110">スーツケースの中を見ましたが 着て行けるようなものがありません</seekvideo>

図 3.3-2　日本語のデータの一例

<seekvideo id="1157">Trong vài tháng vừa qua, tôi đã có những chuyến đi kéo dài nhiều tuần</seekvideo>
<seekvideo id="4532">chỉ với một va li quần áo.</seekvideo>
<seekvideo id="6892">Một ngày, tôi được mời đến một sự kiện quan trọng</seekvideo>
<seekvideo id="9075">và tôi muốn mặc bộ nào đó mới và đặc biệt cho dịp này.</seekvideo>
<seekvideo id="12140">Nên tôi bới tung va li lên nhưng chẳng tìm được bộ nào thích hợp cả</seekvideo>

図 3.3-3　ベトナム語のデータの一例

（２）調査の観点での検討

（２－１）作成可能な対訳コーパスの分量

TED Talks のサイトで公開されている講演データの数は言語ごとに以下のようになっている[35]。

英語　　：2,085
日本語　　：2,001
ベトナム語：1,576
タイ語　　：　995

越日対訳コーパスを作るには、日本語とベトナム語の両方で翻訳されているデータが必要であり、それらは 1,560 件である。またタイ語については、967 件である。これらの講演データを用いて以下の分量の対訳コーパスを作成することができた。

越日対訳コーパス：153,491
泰日対訳コーパス：　96,820

（２－２）作成される対訳コーパスの特徴

TED コーパスの第一の特徴は、講演の書き起こしデータであるため、表現が口語体であることである。そのため、この対訳コーパスを統計翻訳の学習に用いた場合、翻訳結果の文体が口語体となる可能性があり、訳文が文語体で書かれる特許明細書の翻訳としてそぐわないものとなる恐れがある。また（１）で述べたように、日本語のテキストには句点が記載されていない。そのため、本対訳コーパスを用いて学習した統計翻訳システムでは、文末であっても句点のない訳文が生成される可能性がある。

[35] https://wit3.fbk.eu/mono.php?release=XML_releases&tinfo=cleanedhtml_ted

（２－３）作成の具体的手順（作成の準備含む）

a) XML 形式で公開されている各言語の講演データをダウンロードする。(越日対訳コーパスを作る場合には、ベトナム語と日本語だけでなく英語のデータも必要。)

b) 講演データの ID として<url>タグ内にある URL を使用し、ベトナム語と日本語のデータで共通する講演のリストを作成する。

c) (b)で得られた各講演に対し、英語のデータで文頭となるデータの id を求める。図 3.3-1 に示したデータの場合、1 行目の id である 1157 に加え、末尾が文末のピリオドである id が 4532、9075 の次のデータの id である 6892、12140 を求める。

d) ベトナム語と日本語の講演データに対し、(c)で得られた文頭 id から次の文頭 id までのデータを 1 文として抽出し、対訳として出力する。

（２－４）作成に要する期間（作成の準備含む）

XML 形式の講演データは各言語で 10 から 20MB 程度(zip 形式)であるため、データのダウンロードは短時間で終了する。また（２－３）で述べた作業に要する時間も、それぞれの作業に必要なプログラムの開発時間を除けば、30 分程度で実施できる。

（２－５）作成コスト

ダウンロードしたデータを解析して対訳辞書として抽出するプログラムを作成する必要があるが、一般的なプログラミングスキルを有する者が行えば一日程度で作成できるため、考慮すべき作成コストはほとんどないと言える。

（２－６）作成に際して存在する課題

英語以外の言語に翻訳がされている場合であっても、翻訳結果が途中までしか記載されていない場合があり、対訳コーパスに加工する際にそのような事例を削除する必要がある。

３．４　Wikipedia からの対訳コーパス・辞書作成

３．４．１　Wikipedia の記事内容を利用した対訳コーパス・辞書作成

（１）法的・技術的な観点から見た方法の妥当性

Wikipedia のテキストは「クリエイティブ・コモンズ 表示-継承 3.0 非移植[36]」ライセンスの下で利用可能となっている。作成された越日対訳コーパスが統計翻訳の学習目的にのみ用いられ、コーパス自体を配布することがなければ著作権侵害となる恐れはないので、法的な問題はないと考えられる。また当該対訳コーパスを用いて作成される対訳辞書も、元のコンテンツに対して統計的な解析を行って得られるものであって同コンテンツの著作

36 https://ja.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License　(2017 年 1 月 9 日確認)

物性を侵害する内容とはならないので、法的な問題はないと考えられる。

一方技術的には、Wikipedia の記事内容を利用して対訳コーパスを作成するのは必ずしも簡単ではない。単語同士が翻訳関係にある場合であっても、それらの記事内容は必ずしも翻訳関係にはない。すなわち、Wikipedia の記事を行っているボランティアは、それぞれの言語において独自に記事の執筆を行っており、必ずしも最初に英語の記事が書かれ、それが各国語に翻訳されているというわけではない。仮に最初にそのように作成されたとしても、言語ごとに独自に加筆修正が行われるため、初めの段階では文同士の対応関係があったとしても、それが継続的に維持・管理されていることは期待できない。

以上より、Wikipedia の記事内容から越日対訳コーパスを抽出し、該対訳コーパスから辞書を生成する方法については実現可能性が低いと判断し、詳細な検討は省略する。

３．４．２　Wikipedia の見出しを利用した対訳辞書作成

（１）法的・技術的な観点から見た方法の妥当性

Wikipedia のテキストを統計情報の取得に利用する場合と異なり、Wikipedia 上の記事全部の見出しを利用して辞書を作成する場合には、二次的著作物を作成することに該当すると考えられる。そのため、作成した辞書を用いる場合には、「クリエイティブ・コモンズ 表示-継承 3.0 非移植」ライセンスに従い、クレジットを表示するとともに、ライセンス継承の条件に従い、作成した辞書を頒布する必要があると考えられる。

技術的には、言語間リンクで対応が取れている２つの言語での見出しを対の形で抽出して出力する処理を行うだけなので、容易に実施可能である。

（２）調査の観点での検討

（２－１）作成可能な対訳辞書の分量

2016 年 8 月時点のデータを用いて作成された対訳辞書の項目数は以下の通りである。

越日辞書：　16.3 万語

泰日辞書：　 7.8 万語

（２－２）作成される対訳辞書の特徴

辞書の見出しは Wikipedia の記事のタイトルであるため、基本的には名詞である。

（２－３）作成の具体的手順（作成の準備含む）

a) Wikipedia のサイトからデータをダウンロードする。

b) ダウンロードしたデータを解析して対訳辞書として抽出するプログラムを準備する。

c) 抽出処理を行う。

{"type":"item","id":"Q1","labels":{"en":{"language":"en","value":"universe"},
"ja":{"language":"ja","value":"¥u5b87¥u5b99"},
"vi":{"language":"vi","value":"v¥u0169 tr¥u1ee5"},
"th":{"language":"th","value":"¥u0e40¥u0e2d¥u0e01¥u0e20¥u0e1e"},

図 3.4-1 Wikipedia のデータの一例
（日越泰に関連する一部のみを抜粋）

（２－４）作成に要する期間（作成の準備含む）

Wikipedia のサイトからデータをダウンロードし、その後ダウンロードしたデータを解析して対訳辞書として抽出するのに要する時間は半日程度である。

（２－５）作成コスト

ダウンロードしたデータを解析して対訳辞書として抽出するプログラムを作成する必要があるが、専門知識を有する者が行えばほとんどかからない。

（２－６）作成に際して存在する課題

特になし。

３．５　人手翻訳による対訳コーパス・辞書作成

ベトナム語の公報を人手により日本語に翻訳し、ベトナム語公報テキストと日本語テキストから対訳コーパス及び辞書を作成する方法である。

（１）法的・技術的な観点から見た方法の妥当性

インターネット上でアクセス可能な言語資源を用いる場合と違い、独自に翻訳を行うのであるから、法的な問題は存在しない。

対訳コーパスの作成に関する技術的な問題はない。翻訳する際にベトナム語のテキストを文単位で翻訳してもらうことにより、そのまま対訳コーパスとなる。対訳辞書については、できあがった対訳コーパスをもとに、フレーズテーブル作成ツールを用いて辞書登録候補語を抽出して一から辞書を作成するというアプローチも当然可能であるが、日本語の特許で出現頻度の高い単語を決定し、それをベトナム語に翻訳することで基本語レベルの対訳辞書を先に作成することで、より少ないコストで同じ規模の対訳辞書を作成することができると考えられる。

（２）調査の観点での検討

本項目に関しては、本調査事業の仕様書の指示に従い、コストと作成に関する期間のみ

調査する。なお、コストと期間の見積もりにあたり、従来の統計翻訳に関する研究発表などを参考にして、作成する対訳コーパスのサイズを統計翻訳である程度の翻訳精度が期待できる 100 万文対と仮定した。ただし、現在急速に研究が進んでいるニューラルネットワークを用いた機械翻訳技術では、従来の統計翻訳よりも少ない対訳コーパスで同程度の性能を実現できるという研究報告[37]もあることから、実用的な機械翻訳システムを作るのに対訳コーパスとして最低限 100 万文対のコーパスが必要であると言っているわけではないことに留意されたい。

（２－１）作成コスト

英語や中国語のような翻訳需要や翻訳者の多い言語と違い、ベトナム語やタイ語では翻訳者の数や専門分野が限定されることから、これらの言語と比較すると、ベトナム語やタイ語の翻訳はどうしても割高となってしまう。実際に、翻訳会社 3 社に対して特許明細書の翻訳の見積もりを依頼し、100 万文を翻訳するコストを算出したところ以下のようになった。

越日 100 万文の翻訳コスト：　11.7 億円

泰日 100 万文の翻訳コスト：　12.4 億円

なお、上記算出にあたっては、2 章のツール評価で使用したベトナム語 7,473 文とタイ語 3,018 文のテキストから 1 文あたりの平均単語・文字数を求めて利用した。ベトナム語は 1 文平均 35 単語（音節）、タイ語は 1 文平均 73 文字であったので、これにそれぞれ翻訳単価をかけて 1 文平均のコストを出しそれを 100 万倍した。

（２－２）作成に要する期間（作成の準備含む）

（２－１）と同時に翻訳に要する時間の見積もりを依頼し、100 万文を翻訳する期間を算出したところ以下のようになった[38]。

越日 100 万文の翻訳期間：　119 人年　（40 人体制で約 3 年）

泰日 100 万文の翻訳期間：　133 人年　（40 人体制で約 3.3 年）

なお、上記算出にあたっては、翻訳会社に見積もりを依頼した際のテキストのワード・文字数とその翻訳に要すると回答した期間を元に 100 万文のテキストを翻訳するための所要期間を算出した。ただし（２－１）でも述べたように、当該言語では翻訳者の数が英語や中国語に比べて少ないため、これだけの規模の翻訳者が同時に確保できるかは不明である。

---

[37] Katsuhito Sudoh and Masaaki Nagata: Chinese-to-Japanese Patent Machine Translation based on Syntactic Pre-ordering for WAT 2016.
http://lotus.kuee.kyoto-u.ac.jp/WAT/papers/WAT2016-proceedings.pdf （最終検索日：2017 年 2 月 17 日）

[38] 翻訳期間は 2 社からしか見積もりが得られなかったため、2 社の平均により算出した。

３．６　他事業の報告書に記載された言語資源を利用した対訳コーパス・辞書作成

特許庁の他の調査事業においてベトナム語やタイ語の言語資源として判明しているものの中で、商用利用の可能性があり、ある程度の規模の対訳コーパスが作成できそうなデータにOPUS(Open Parallel Corpus)のサイト[39]で公開されている言語資源がある。本調査で対象としているベトナム語と日本語のデータの中で大規模なデータが存在するのは、主に映画等のサブタイトルのデータとソフトウェアのマニュアル等のデータの 2 種類であるが、サブタイトルのデータは元の作品の著作権が関係してくることと、TED 以上に口語体で、しかも日常的な崩れた表現となっている可能性が高く、特許用 SMT の学習に用いるのは不適当だと思われる。一方ソフトウェアのデータであれば、SMT の学習に使う場合には、著作権の問題もないと思われるので、後者に関して調査を行った。具体的には GNOME というソフトウェアに関するデータである。

（１）法的・技術的な観点から見た方法の妥当性

既にこれまで取り上げてきた言語資源と同様、OPUS のサイトのデータについても、それらを利用して作成した対訳データを SMT の学習に使うことについての法的な問題はないと考えられる。

GNOME のデータはソフトウェアのモジュールごとにドキュメントがあり、その単位で対応づけできるため、技術的には、本データを用いた対訳コーパス作成は、パテントファミリーとして対応付けられた特許公報を用いた対訳コーパス作成とほぼ同じやり方で行うことができ、技術的な問題もない。以下にサンプルを示すように、各文書は XML 形式で公開されているが、文単位での対応は取られていないため、本調査で公報データを用いて行っている文アライメントツールでのアライメント処理が必要となる。

[39] http://opus.lingfil.uu.se/ （最終検索日：2017 年 2 月 20 日）

```
<?xml version="1.0" encoding="utf-8"?>
<cesDoc type="text" version="1.0" TEIform="TEI.2">
  <cesHeader type="text" creator="joerg" date.created="Sun Aug 17 17:38:40 2014">
        （中略）
  </cesHeader>
  <text>
    <body>
      <div type="section">
<s id="s1">Chương trình này là phần mềm tự do; bạn có thể phát hành lại nó và/hoặc sửa đổi nó với điều kiện của Giấy Phép Công Cộng GNU (LGPL) được xuất bản bởi Tổ chức Phần mềm Tự do; hoặc phiên bản 2 của Giấy Phép này, hoặc (tùy chọn) bất kỳ phiên bản mới hơn.</s>
```

図 3.6-1　ベトナム語データの一例

```
<?xml version="1.0" encoding="utf-8"?>
<cesDoc type="text" version="1.0" TEIform="TEI.2">
  <cesHeader type="text" creator="joerg" date.created="Sat Aug 16 16:15:44 2014">
          （中略）
    </cesHeader>
    <text>
      <body>
        <div type="section">
<s id="s1">このプログラムはフリーソフトウェアです。あなたはこれを、フリーソフトウェア財団によって発行された GNU 劣等一般公衆利用許諾契約書(バージョン 2 か、希望によってはそれ以降のバージョンのうちどれか)の定める条件の下で再頒布または改変することができます。</s>
```

図 3.6-2　日本語データの一例

（2）調査の観点での検討

（2－1）作成可能な対訳コーパスの分量

ベトナム語と日本語の両方の訳がある文書は 1,187 件であり、それらを利用して文アライメントを行った場合、理想的な対応ができた場合には約 60 万文対の対訳が作成可能であることが分かった。同様にタイ語と日本語では、1,093 件の文書が共通しており、理想的な

対応で得られる対訳数は約 53 万文対である。

（２－２）作成される対訳コーパスの特徴

本データに収録されているテキストは、主にソフトウェアで表示されるメッセージのテキストやメニューの項目である。理解のしやすさを念頭に下に英語も含めて 4 言語で対応する文の例を示したが、メニュー項目は日本語では主語のない平叙文であるが、英語では命令文となっている。これはベトナム語やタイ語でも同様である。したがって、このような対訳を用いて学習した SMT では、主語のない日本語平叙文が他の言語では命令文として訳出される可能性もあるが、日本語への翻訳方向の場合には影響は軽微であると考えられる。

日本語：	ウィンドウのアイコンを指定する
英語：	Set the window icon
ベトナム語：	Lập biểu tượng cửa sổ
タイ語：	กำหนดไอคอนของหน้าต่าง

図 3.6-3　各言語で対応するテキストの一例

（２－３）作成の具体的手順（作成の準備含む）

a) XML 形式で公開されている各言語のデータをダウンロードし解凍する。
b) 図 3.6-1 と 2 の例に示す通り、各テキストは タグの中にあり、文ごとに<s>タグがついているので、<s>タグごとに 1 文として抽出することで文分割されたテキストが作成できる。
これ以降は、パテントファミリーを用いた場合の越日対訳コーパスを作成した手順と同じであるので、詳細な説明は省略する。
c) ベトナム語のテキストを英語に翻訳する。
d) 英語に翻訳したテキストと日本語テキストを、文アライメントツールを用いて対応付ける。
e) 英語テキストを介して、元のベトナム語の文と日本語の対応付けを行う。

（２－４）作成に要する期間（作成の準備含む）

データのダウンロードは、一般的なインターネット環境であれば、各言語 10 分程度である。また文アライメントの準備として、ベトナム語のテキストを英語に翻訳する必要がある。インターネット上の翻訳サービスを API を使わずに手作業で利用して翻訳する場合、ベトナム語文が計 7,540 文からなる公報データ 30 文書を翻訳するのに約 45 分であったので、50 万文では約 50 時間程度必要となる。また、その後の文アライメント処理は、使用す

る計算機の演算性能にもよるが、調査者の環境では、ベトナム語文が計 7,540 文からなる公報データ 30 文書の対応付けに約 13 分かかったので、50 万文では 15 時間程度と見積もられる。

（２－５）作成コスト

50 万文のベトナム語テキストを翻訳する作業の人件費が若干発生するが、作業時間が 50 時間程度とそれほど大きなコストではない。また文アライメントツールとして内山らのツールを使う場合には、その導入費用が必要となるが、もともとパテントファミリーデータを用いた対訳作成において導入するのが前提であれば追加コストは発生しない。

（２－６）作成に際して存在する課題

本方式固有の課題は特に存在しない。

３．７　その他の言語資源を利用した対訳コーパス・辞書作成

３．７．１　PCT 条約を用いた対訳コーパスの作成

（１）法的・技術的な観点から見た方法の妥当性

元の英語の条文を各国語に翻訳されたテキストの翻訳権は、翻訳を行った者が有するが、本データを用いた対訳コーパスを SMT の学習目的のみに使用する場合には、これまで述べてきたように法的な問題はないと考えられる。

また技術的にも、条文は細かく章立てされており、文の単位も概ね英語原文の単位が保存されている場合が多いことから、対応付けは容易である。

（２）調査の観点での検討

（２－１）作成可能な対訳コーパスの分量

日本語に翻訳された条文は約 570 文であるが、インターネット上に公開されていたベトナム語の条文は翻訳が不完全なため、得られた対訳は約 180 文対に止まっている。一方、タイ語は完全な翻訳があったため条文の最後まで対応付けができ約 570 文対のコーパスが作成できた。

（２－２）作成される対訳コーパスの特徴

技術文献ではないため、特許翻訳用の SMT の学習に利用しても導入効果は大きくない可能性がある。

（２－３）作成の具体的手順（作成の準備含む）

ベトナム語条文と日本語条文をもとに手作業にて対応付けを行った。

（２－４）作成に要する期間（作成の準備含む）

2 日程度。

（２－５）作成コスト

対応付けを行う作業を実施する作業者の人件費。

（２－６）作成に際して存在する課題

特になし。

３．７．２　IPC 分類を用いた対訳コーパスの作成

特許の分類体系の一つである IPC 分類は、「A01B 1/14」のようなコードで特許文献に記載された技術内容を体系化するためのものであり、各コードにはそのコードに対する説明文が付与されている。したがって、コードをピボットとして結びつけることで、対訳コーパスを作ることができる。

（１）法的・技術的な観点から見た方法の妥当性

各国語に翻訳した分類表全体の翻訳権は翻訳を行った者が有するが、コードに付与された説明文から作成した対訳コーパスを SMT の学習目的のみに使用する場合には、これまで述べてきたように法的な問題はないと考えられる。

IPC コードを手掛かりとして２つの言語の説明文の対応を取ればよいので、２つの言語の分類表から対訳コーパスを作成する際の技術的課題はない。

（２）調査の観点での検討

（２－１）作成可能な対訳コーパスの分量

IPC 分類表は毎年のように変更が加えられるため、どの時点のバージョンを翻訳したのかにより対応付け可能なコードの数が変化するが、2017 年 2 月現在日本国特許庁のサイトで公開されている日本語版の分類表における IPC コードの数で見積もると約 74,000 文対の対訳が作成可能である。

（２－２）作成される対訳コーパスの特徴

各 IPC コードに対する説明文は、日本語版を見る限り基本的にほとんどが名詞句である。

（２－３）作成の具体的手順（作成の準備含む）

①日本語版は Excel ファイルのものが公開されているので、それを TSV 形式で出力し、「記号」列が IPC コードであるデータを抽出する。

②ベトナム語版を OCR でテキスト化し、IPC コードの説明文を抽出する[40]。

[40] 本来は pdf ファイルから直接テキストデータを取り出せるが、現在公開されている pdf

③上の①②で作成したテキストから共通する IPC コードを手掛かりに文アライメントを行う。

（２－４）作成に要する期間（作成の準備含む）

ここで検討した手法において最も時間を有するのはベトナム語分類表の OCR によるテキスト化作業である。トライアルとしてベトナム語版の分類表 3 ページを OCR で読み取り、読み取り結果のチェック・修正作業を行ったところ、1 ページの作業に要する時間は、約 10 分であった。IPC 分類表はトータルで 3,125 ページあるため、チェック作業に要する時間は 74 日程度必要である[41]。

（２－５）作成コスト

OCR を用いた分類表のテキスト化を行う作業を実施する作業者の人件費。

（２－６）作成に際して存在する課題

特になし。

ファイルに埋め込まれているテキストを Acrobat Reader 等のビューワで表示しテキストをコピーペーストで取り出しても、ビューワで表示されるテキストとは異なるテキストしか抽出できない。そのため、そのままでは利用することができず、OCR を用いてテキスト化する必要があると判断した。ただしベトナム国家知的財産庁から直接電子化データを入手できればもちろんこの作業は省略でき、より短時間低コストでの作成が可能となる。

[41] 一日の作業時間は 420 分で計算した。

付録

付録１ 対訳コーパス

（１）越日対訳コーパス（一部）

#	ベトナム語	日本語
1	Lĩnh vực kỹ thuật được đề cập	技術分野
2	Sáng chế đề cập tới vật ghi đĩa quang trong đó dữ liệu chính được ghi dưới dạng kết hợp các lõm và các vùng dẫn được tạo ra trên một mặt của nền, lớp phản xạ và lớp phủ được chồng lên trên phía nền nơi mà các lõm và vùng dẫn được tạo ra, dữ liệu phụ được ghi dưới dạng các dấu được tạo ra bằng cách chiếu ánh sáng laze có công suất ghi lên lớp phản xạ và mức đầu ra của tín hiệu đọc sẽ được tăng tại các phần mà các dấu được tạo ra, thiết bị và phương pháp đọc dùng cho vật ghi đĩa quang và thiết bị và phương pháp ghi dùng để ghi dữ liệu phụ lên vật ghi đĩa quang.	[0001] 本発明は、基板の一の面にピット及びランドの組み合わせによって主データを記録し、基板のピット及びランドが形成された面を覆って反射膜とカバー層が積層され、記録パワーによるレーザ光の反射膜への照射により形成されるマークによって副データが記録される光ディスク記録媒体であって、マークが形成された部分での再生信号の出力レベルが上昇するように構成されている光ディスク記録媒体の再生を行う再生装置及び再生方法に関し、さらには、光ディスク記録媒体に副データを記録するための記録装置及び記録方法に関する。
3	Tình trạng kỹ thuật của sáng chế	背景技術
4	Các đĩa quang được sử dụng để ghi thông tin bao gồm ROM (read-only memory - bộ nhớ chỉ đọc).	[0002] 情報記録媒体として用いられる光ディスクとして、再生専用型の光ディスクである R OM (Read- Only Memory)ディスクがある。
5	Đĩa ROM được sử dụng rộng rãi trên thế giới như là bộ đĩa do nhiều nền sao của nó có thể được sản xuất trong một thời gian ngắn bằng cách đúc phun chất dẻo bằng khuôn dập có các lõm và vùng dẫn được tạo ra trên đó từ trước.	この ROMディスクは、予めピットやランドが形成されたスタンパを装着した金型装置を用いてプラスチックを射出成型することにより、短時間で大量のレプリカ基板を製造可能であることからパッケージメディアとして広く利用されている。
6	Trong số các đĩa ROM loại này, CD (Compact Disk - Đĩa compact) và DVD (Digital Versatile Disk - Đĩa đa năng số), chẳng hạn, được sử dụng rộng rãi như là các vật ghi dùng để ghi thông tin nội dung như âm nhạc, video, V.V..	この種の ROMディスクのうち、例えば CD (Compact Disc) や DVD (Digital Versatile Disc)は、音楽や映像等のコンテンツ情報を記録する記録媒体として広く用いられて!/ヽる。
7	Các đĩa trên đó có sao chép trái phép dữ liệu được ghi trong đĩa ROM được bán như là bộ đĩa, được gọi là đĩa giả, đã được sản xuất và gây tổn hại đối với các quyền lợi của người có bản quyền hợp pháp đối với dữ liệu trong đĩa ROM.	従来、ノッケージメディアとして販売されて、る ROMディスクを基にその記録データを違法複写したヽゎゆる偽造ディスクが作成され、正規の著作権を有する者の利益が侵害されて、ると、う問題がある。
8	Thông thường, các đĩa giả được sản xuất bằng cách tạo khuôn dập bằng cách tạo bản gốc dựa vào các tín hiệu đọc được từ đĩa được xác nhận và sao chép các đĩa quang bằng khuôn dập, hoặc bằng cách sao chép các tín hiệu được đọc từ đĩa được xác nhận lên các đĩa có thể ghi được.	一般的に、偽造ディスクは、正規の光ディスク力再生した信号を基にマスタリング工程によりスタンパを作成し、このスタンパを元にして複製ディスクを製造、又は、正規の光ディスク力再生した信号を記録可能なディスクに複写することで作成される。
9	Cho đến nay, nhiều kỹ thuật ngăn ngừa sao chép khác nhau đã được đề xuất nhằm ngăn không cho các vật ghi thông tin giả như vậy được sản xuất bởi những người không có quyền hợp pháp.	正当な権限を有することなく作成される偽造の情報記録媒体の製造を防止するため、種々のコピー防止技術が提案されている。
10	Một trong số các kỹ thuật như vậy là gắn thêm thông tin nhận dạng duy nhất vào từng đĩa chẳng hạn.	その 1つとして、例えばディスクごとに異なる識別情報を付加する技術が知られている。

（２）越英対訳コーパス（一部）

#	ベトナム語	英語
1	Lĩnh vực kỹ thuật được đề cập	FIELD OF THE INVENTION
2	Sáng chế đề cập tới vật ghi đĩa quang trong đó dữ liệu chính được ghi dưới dạng kết hợp các lõm và các vùng dẫn được tạo ra trên một mặt của nền, lớp phản xạ và lớp phủ được chồng lên trên phía nền nơi mà các lõm và vùng dẫn được tạo ra, dữ liệu phụ được ghi dưới dạng các dấu được tạo ra bằng cách chiếu ánh sáng laze có công suất ghi lên lớp phản xạ và mức đầu ra của tín hiệu đọc sẽ được tăng tại các phần mà các dấu được tạo ra, thiết bị và phương pháp đọc dùng cho vật ghi đĩa quang và thiết bị và phương pháp ghi dùng để ghi dữ liệu phụ lên vật ghi đĩa quang.	The present invention relates to an optical-disk recording medium in which main data is recorded in the form of a combination of pits and lands formed on one side of a substrate, a reflective layer and cover layer are stacked over the substrate side where the pits and lands are formed, sub data is recorded in the form of marks formed by irradiating laser light having a writing power to the reflective layer and the output level of a reading signal will be raised at the portions where the marks are formed, a playing apparatus and method for the optical-disk recording medium and a recording apparatus and method for recording sub data to the optical-disk recording medium.
3	Tình trạng kỹ thuật của sáng chế	BACKGROUND ART
4	Các đĩa quang được sử dụng để ghi thông tin bao gồm ROM (read-only memory - bộ nhớ chỉ đọc).	The optical disks used to record information include the ROM (read-only memory).
5	Đĩa ROM được sử dụng rộng rãi trên thế giới như là bộ đĩa do nhiều nền sao của nó có thể được sản xuất trong một thời gian ngắn bằng cách đúc phun chất dẻo bằng khuôn dập có các lõm và vùng dẫn được tạo ra trên đó từ trước.	The ROM disk is widely used as a package medium over the world because many replica substrates thereof can be produced in a short time by injection molding of plastics with a stamper having pits and lands formed thereon in advance.
6	Trong số các đĩa ROM loại này, CD (Compact Disk - Đĩa compact) và DVD (Digital Versatile Disk - Đĩa đa năng số), chẳng hạn, được sử dụng rộng rãi như là các vật ghi dùng để ghi thông tin nội dung như âm nhạc, video, V.V..	Of the ROM disks of this type, CD (Compact Disk) and DVD (Digital Versatile Disk), for example, are widely used as recording media to record content information such as music, video, etc.
7	Các đĩa trên đó có sao chép trái phép dữ liệu được ghi trong đĩa ROM được bán như là bộ đĩa, được gọi là đĩa giả, đã được sản xuất và gây tổn hại đối với các quyền lợi của người có bản quyền hợp pháp đối với dữ liệu trong đĩa ROM.	Disks having illegally copied thereto data recorded in a ROM disk sold as a package medium, so-called counterfeit disks, have ever been produced and prejudicial to the interests of a person having the regular copyright for the data in the ROM disk.
8	Thông thường, các đĩa giả được sản xuất bằng cách tạo khuôn dập bằng cách tạo bản gốc dựa vào các tín hiệu đọc được từ đĩa được xác nhận và sao chép các đĩa quang bằng khuôn dập, hoặc bằng cách sao chép các tín hiệu được đọc từ đĩa được xác nhận lên các đĩa có thể ghi được.	Generally, the counterfeit disks are produced by forming a stamper by mastering on the basis of signals read from an authenticated disk and replicating optical disks by the stamper, or by copying signals read from the authenticated disk to recordable disks.
9	Cho đến nay, nhiều kỹ thuật ngăn ngừa sao chép khác nhau đã được đề xuất nhằm ngăn không cho các vật ghi thông tin giả như vậy được sản xuất bởi những người không có quyền hợp pháp.	Various techniques for copy prevention have been proposed heretofore to prevent such counterfeit information-recording media from being produced by those having no due right.
10	Một trong số các kỹ thuật như vậy là gắn thêm thông tin nhận dạng duy nhất vào từng đĩa chẳng hạn.	One of such techniques is to append, for example, unique identification information to each of disks.
11	Có thể xây dựng hệ thống trong đó thông tin nhận dạng duy nhất được gắn vào từng đĩa bằng kỹ thuật này, và máy đọc đĩa đọc thông tin nhận dạng và gửi thông tin này tới máy tính phục vụ bên ngoài thông qua mạng.	There can be built a system in which unique identification information is appended to each disk with this technique, and a disk player reads the identification information and sends it to an external server via a network.
12	Thậm chí nếu các đĩa giả như vậy đã được sản xuất và phân phối, thì máy tính phục vụ bên ngoài sẽ phát hiện một lượng lớn cùng thông tin nhận dạng và do đó hệ thống có thể phát hiện ra là các đĩa giả đã được sản xuất và phân phối như vậy.	Even if such counterfeit disks have been produced and distributed, the external server will detect a large amount of the same identification information and the system can thus detect that the counterfeit disks have been so produced and distributed.
13	Hơn nữa, hệ thống còn có thể nhận dạng người sản xuất hay phân phối đĩa giả bằng cách nhận dạng máy đọc đĩa đã gửi thông tin nhận dạng phát hiện được tới máy tính phục vụ bên ngoài.	Further, the system can also identify a counterfeit disk maker or distributor by identifying a disk player having sent the detected identification information to the external server.
14	Ngay cả thông tin nhận dạng duy nhất đối với từng đĩa được xác nhận cần được ghi để không dễ dàng bị sao chép như nêu trên bằng ổ đĩa sẵn có trên thị trường, điều này sẽ có tác dụng bảo vệ bản quyền đối với dữ liệu chính trong đĩa.	Even identification information unique to each authenticated disk should be recorded not to easily be copied as above by a commercially available disk drive, which will be useful for protection of the copyright for the main data in the disk.

（３）泰日対訳コーパス（一部）

#	タイ語	日本語
1	ข้อต่อที่มีความเร็วคงที่ประเภทพลันจิ้ง	【公報】しゅう動式の等速ジョイント
2	สาขาวิทยาการที่เกี่ยวข้องกับการประดิษฐ์	【技術分野】
3	วิศวกรรมศาสตร์ที่ซึ่งการประดิษฐ์นี้จะเกี่ยวข้องกับข้อต่อของยานพาหนะ เช่น ข้อต่อที่มีความเร็วคงที่ประเภทพลันจิ้ง (a plunging type constant velocity joint)	【０００１】本願は、しゅう動式の等速ジョイント等の車両のジョイントに関する。
4	ภูมิหลังของศิลปะหรือวิทยาการที่เกี่ยวข้อง	【背景技術】
5	ดามที่แสดงเอาไว้ในรูปที่ 1A และ 1B ข้อต่อแบบดับเบิ้ลออฟเซ็ตดั้งเดิม 10 ประกอบไปด้วยส่วนข้อต่อภายนอก 12 ส่วนข้อต่อภายใน 14 กรง (a cage) 16 และลูปปืนหนึ่งเม็ดเป็นอย่างน้อย 18	【０００２】図１の（Ａ）および（Ｂ）に示すように、従来のダブルオフセット型のジョイント１０は、外側ジョイント部１２と、内側ジョイント部１４と、ケージ１６と、少なくとも１つのボール１８とを含んでいる。
6	ดามปกติ ลูกปืนจำนวนหนึ่ง 18 ถูกจัดให้มีเอาไว้	通常は複数のボール１８が提供される。
7	ส่วนข้อต่อภายนอก 10 มีรูปร่างทรงกระบอกที่ถูกสร้างขึ้นพร้อมกับรางนำทางเส้นตรงที่มีพื้นที่ว่างตามเส้นรอบวงจำนวนหนึ่ง 20 ที่มีส่วนโค้งโกธิก (a gothic arc) หรือรูปวงรี	外側ジョイント部１２は、ゴシックアーチ状または楕円形状の複数の円周方向に隔てられた線形のガイドトラック２０を形成され、円筒形状をもつ。
8	ส่วนข้อต่อภายใน 14 มีรางนำทางเส้นตรงที่มีพื้นที่ว่างเส้นรอบวง 22 ที่ถูกสร้างขึ้นบนพื้นผิวทรงกลมภายนอก 24	内側ジョイント部１４は、外部球面２４に形成された、円周方向に隔てられた線形のガイドトラック２２を有している。
9	กรง 6 กั้นลูกปืนจำนวนหนึ่ง 18 ในช่องจำนวนหนึ่ง 26 ที่เว้นระยะห่างตามเส้นรอบวงรอบกรง 16	ケージ１６は、ケージ１６の周りに円周方向に隔てられた複数のポケット２６に複数のボール１８を保持している。
10	กรง 16 มีพื้นผิวทรงกลมที่เว้าเข้าด้านในที่ R2 และพื้นผิวทรงกลมที่นูนออกด้านนอกที่ R1	ケージ１６は、Ｒ２の内側の凹状の球面を有し、Ｒ１の外側の凸状の球面を有している。
11	R1 และ R2 ถูกเอียงได้โดย e ไปยังด้านที่อยู่ตรงกันข้ามของจุด O ไปยังจุด O1, O2 /// ในทิศทางตามแนวแกนจากศูนย์กลางของช่องลูกปืนในที่ซึ่งพื้นผิวทรงกลมที่นูนออกด้านนอกที่ R1 สัมผัสกับขนาดกระบอกสูบ 28 ของส่วนข้อต่อภายนอก 12	Ｒ１とＲ２とは、ボールポケットの中心から軸方向にある、Ｒ１の外側の凸状の球面が外側ジョイント部１２の円筒状のボア２８と接触する点Ｏ１、Ｏ２に対して点Ｏとは反対の側からｅ分、オフセットされている。
12	พื้นผิวทรงกลมที่เว้าเข้าด้านในของกรง 16 ที่ R2 สัมผัสกับพื้นผิวทรงกลมที่นูนออกด้านนอก 24 ของส่วนข้อต่อภายใน 14	Ｒ２におけるケージ１６の内側の凹状の球面は、内側ジョイント部１４の外側の凸状の球面２４と接触している。
13	ส่วนข้อต่อภายใน 14 ยังคงมีพื้นผิวภายใน 30 สำหรับการต่อเข้ากับเพลา (ไม่ได้แสดงเอาไว้)	内側ジョイント部１４は、シャフト（不図示）と接触するための内表面３０を有している。
14	ในโครงสร้างของข้อต่อดังกล่าว ถ้าแรงบิดที่เหมาะสมถูกกระทำกับข้อต่อ 10 แรงจะกระทำต่อร่องลูกปืน (ball track) 20, 22 หรือลูกปืน 18 ในทิศทางที่ตั้งฉากกับร่องลูกปืน 20, 22	【０００３】このようなジョイント構造において、一定のトルクがジョイント１０にかかると、ボールトラック２０、２２、またはボール１８に、ボールトラック２０、２２の法線方向に負荷がかかる。
15	แรงอีกแรงหนึ่งที่ได้รับเพียงบางส่วนมาจากแรงบนร่องลูกปืน 20, 22 ที่กระทำบนพื้นผิว 32, 34 ของช่องลูกปืน 26 ของกรงในทิศทางตามแนวแกน Z ที่มุมในการต่อกัน	ボールトラック２０、２２に対する負荷に一部起因する別の負荷が、ケージのボールポケット２６の表面３２、３４に、軸方向Ｚの関節角（articulation angle）でかかる。
16	ในสภาพนี้ ลูกปืน 18 สัมผัสกับร่อง 20, 22 ที่มุมกด A	この条件において、ボール１８がトラック２０、２２に圧力角Ａで接触する。
17	พื้นที่สัมผัส CA ถูกสร้างขึ้น ที่ทำให้เกิดรูปแบบของวงรีที่ถูกจำกัดขอบเขตได้โดยความยาวที่ยาวขึ้นของหน้าสัมผัสวงรี a ระหว่างลูกปืนนูน 18 และร่องเว้า 22 และความยาวที่สั้นลงของหน้าสัมผัสรูปวงรี b ที่สร้างขึ้นระหว่างลูกปืนนูนและร่องที่ทำเป็นรูปทรงกระบอก 20	凸ボール１８と凹トラック２２との間の楕円接触部の長軸ａと、凸ボールと円筒状トラック２０との間の楕円接触部の短軸ｂにより定義される楕円の形状をとる接触領域ＣＡが生成される。

（4）泰英対訳コーパス（一部）

#	タイ語	英語
1	ข้อต่อที่มีความเร็วคงที่ประเภทพลันจิ้ง	PLUNGING TYPE CONSTANT VELOCITY JOINT
2	สาขาวิทยาการที่เกี่ยวข้องกับการประดิษฐ์	FIELD OF INVENTION
3	วิศวกรรมศาสตร์ที่ซึ่งการประดิษฐ์นี้จะเกี่ยวข้องกับข้อต่อของยานพาหนะ เช่น ข้อต่อที่มีความเร็วคงที่ประเภทพลันจิ้ง (a plunging type constant velocity joint)	The present invention relates to a vehicle joint, such as a plunging type constant velocity joint.
4	ภูมิหลังของศิลปะ	BACKGROUND OF THE INVENTION
5	ตามที่แสดงเอาไว้ในรูปที่ 1A และ 1B ข้อต่อแบบดับเบิ้ลออฟเซ็ตดั้งเดิม 10 ประกอบไปด้วยส่วนข้อต่อภายนอก 12 ส่วนข้อต่อภายใน 14 กรง (a cage) 16 และลูปปืนหนึ่งเม็ดเป็นอย่างน้อย 18	As shown in Figs. 1A and 1 B, a conventional double offset joint 10 comprises an outer joint part 12, an inner joint part 14, a cage 16 and at least one ball 18.
6	ตามปกติ ลูกปืนจำนวนหนึ่ง 18 ถูกจัดให้มีเอาไว้	Typically, a plurality of balls 18 are provided.
7	ส่วนข้อต่อภายนอก 10 มีรูปร่างทรงกระบอกที่ถูกสร้างขึ้นพร้อมกับรางนำทางเส้นตรงที่มีพื้นที่ว่างตามเส้นรอบวงจำนวนหนึ่ง 20 ที่มีส่วนโค้งโกธิก (a gothic arc) หรือรูปวงรี	The outer joint part 10 has a cylindrical shape formed with a plurality of circumferentially spaced linear guide tracks 20 having a gothic arc or an elliptical form.
8	ส่วนข้อต่อภายใน 14 มีรางนำทางเส้นตรงที่มีพื้นที่ว่างเส้นรอบวง 22 ที่ถูกสร้างขึ้นบนพื้นผิวทรงกลมภายนอก 24	The inner joint part 14 has circumferentially spaced linear guide tracks 22 formed on an outer spherical surface 24.
9	กรง 6 กั้นลูกปืนจำนวนหนึ่ง 18 ในช่องจำนวนหนึ่ง 26 ที่เว้นระยะห่างตามเส้นรอบวงรอบกรง 16	The cage 6 retains the plurality of balls 18 in a plurality of pockets 26 circumferentially spaced about the cage 16.
10	กรง 16 มีพื้นผิวทรงกลมที่เว้าเข้าด้านในที่ R2 และพื้นผิวทรงกลมที่นูนออกด้านนอกที่ R1	The cage 16 has an inner concave spherical surface at R2 and an outer convex spherical surface at R1.
11	R1 และ R2 ถูกเอียงได้โดย e ไปยังด้านที่อยู่ตรงกันข้ามของจุด O ไปยังจุด O1, O2 /// ในทิศทางตามแนวแกนจากศูนย์กลางของช่องลูกปืนในที่ซึ่งพื้นผิวทรงกลมที่นูนออกด้านนอกที่ R1 สัมผัสกับขนาดกระบอกสูบ 28 ของส่วนข้อต่อภายนอก 12	R1 and R2 are offset by e to the opposite sides of point O to points 01 , 02 in the axial direction from the center of the ball pocket in which the outer convex spherical surface at R1 contacts a cylindrical bore 28 of the outer joint part 12.
12	พื้นผิวทรงกลมที่เว้าเข้าด้านในของกรง 16 ที่ R2 สัมผัสกับพื้นผิวทรงกลมที่นูนออกด้านนอก 24 ของส่วนข้อต่อภายใน 14	The inner concave spherical surface of the cage 16 at R2 contacts the outer convex spherical surface 24 of inner joint part 14.
13	ส่วนข้อต่อภายใน 14 ยังคงมีพื้นผิวภายใน 30 สำหรับการต่อเข้ากับเพลา (ไม่ได้แสดงเอาไว้)	The inner joint part 14 also has an inner surface 30 for connection with a shaft (not shown).
14	ในโครงสร้างของข้อต่อดังกล่าว ถ้าแรงบิดที่เหมาะสมถูกกระทำกับข้อต่อ 10 แรงจะกระทำต่อร่องลูกปืน (ball track) 20, 22 หรือลูกปืน 18 ในทิศทางที่ตั้งฉากกับร่องลูกปืน 20, 22	In such a joint construction, if a certain torque is applied to the joint 10, a load acts on ball track 20, 22 or balls 18 in the direction normal to the ball track 20, 22.
15	แรงอีกแรงหนึ่งที่ได้รับเพียงบางส่วนมาจากแรงบนร่องลูกปืน 20, 22 ที่กระทำบนพื้นผิว 32, 34 ของช่องลูกปืน 26 ของกรงในทิศทางตามแนวแกน Z ที่มุมในการต่อกัน	Another load derived partly from the load on the ball track 20, 22 acts on surfaces 32, 34 of the ball pockets 26 of the cage in axial direction Z, at an articulation angle.

付録２ 対訳辞書

（１）日越対訳辞書（一部）

#	日本語	ベトナム語１	ベトナム語２	ベトナム語３
1	繊維	từ sợi	các sợi	đều sợi
2	表面	phía bềmặt	bềmặt bằng	được phủ bởi bềmặt
3	データ	dữliệu	dữliệu sẽ	để tập dữliệu
4	シリカ	silic dioxit	bằng silic oxit	đốivới silic oxit
5	量	đạilượng	lượng nằm	đốivới lượng
6	図	các hìnhvẽ từ Fig	trên FIG	ống nano thểhiện
7	重量	theo khốilượng	khốilượng	xácđịnh trọnglượng
8	データレート	Tốcđộ dữliệu	tốcđộ dữliệu	và tốcđộ
9	発明	Phátminh	Sángchế	ưutiên của
10	酸	Axit	axit được	và axit
11	値	trịsố	giátrị	đọc có giátrị
12	化	hóa được	hóa	· hóa
13	水	Nước	nước	đồng sôi
14	樹脂	nhựa	nhựa và	với nhựa
15	炭素	cacbon	Cacbon	than
16	方向	chiều	dọc	đối hướng
17	式	Kiểu	đặcsắc	có côngthức
18	温度	nhiệtđộ là	nhiệtđộ	ở nhiệtđộ
19	実施	trong Vídụ	thểhiện theo các	ưutiên
20	物質	vậtchất	ở đó chất	từ đó chất
21	構成	cấuhình	được hệthống thiếtbị	đứtquãng
22	方法	pháp	và phương pháp	đềxuất các phươngpháp
23	システム	Hệthống	hệthống	· auđio và hệthống
24	金属	quátrình kimloại	kimloại	tạpchất kimloại
25	端子	điệncực	thôngqua cực	đấu dâyđiện
26	形態	một dạng	hìnhthái	đệm có dạng
27	ディスク	đĩa	các đĩa	đĩaquang được
28	構造	cấutrúc	Cấutrúc	độ khuếchđại của cấutrúc
29	レート	tốcđộ	tăng tốcđộ	ở tốcđộ
30	信号	tínhiệu	Tínhiệu	đường truyền tínhiệu

（２）英越対訳辞書（一部）

#	英語	ベトナム語1	ベトナム語2	ベトナム語3
1	example	ví dụ về	ví dụ	làm ví dụ dưới
2	surface	bề mặt càng lớn	bề mặt càng	ở mặt
3	data	dữ liệu	dữ liệu cụ thể	truyền dữ liệu
4	invention	ai 2003	một sáng chế	được một sáng chế
5	present invention	khác sáng chế	et ai 2003	theo sáng chế
6	fiber	bằng sợi	nóng sợi	đổi
7	metal	bằng kim loại	kim loại hiện	vảy
8	temperature	nhiệt độ	nhiệt độ là	ở nhiệt độ
9	amount	là lượng	về lượng	hơn là lượng
10	method	Phương pháp	còn đề xuất phương pháp	đề xuất phương pháp
11	system	Hệ thống	hệ thống	với các hệ thống
12	weight	Trọng lượng	có trọng lượng	đối trọng
13	carbon	than được	than được xử lý	trong cacbon
14	sub data	dữ liệu phụ	dữ liệu phụ có	
15	rate	Tốc độ	cho tốc độ	tỷ lệ
16	access terminal	đầu cuối truy nhập có thể tạo	đầu cuối truy nhập có thể tạo ra	ở đầu cuối truy nhập
17	range	982 đến 1093° C trong khoảng	1093° C trong khoảng	độ nằm
18	side	bên	ở bên trái của nửa puli	ở bên
19	process	quy trình theo sáng chế	quy trình theo	xử lý
20	portion	phần	có phần	ở phía
21	microporous materi	vi xốp vật liệu vi xốp	vi xốp vật liệu vi xốp được	vật liệu vi xốp
22	end	kết thúc	đầu	chỉ đầu
23	water	nước	các dòng nước	đường xả nước
24	silica	silic oxit	silic	đặc silic oxit
25	acid	axit	dung dịch axit	đioxoborinan-5-carboxylic
26	glass substrate	Đế thủy tinh	để thủy tinh	đế thủy tinh
27	case	trường hợp	ngăn	đường tâm L
28	material	chất liệu	chất	xốp
29	acid gas	khí axit	của khí axit sẽ	khí axit sẽ
30	step	bước		

（３）日泰対訳辞書（一部）

#	日本語	タイ語1	タイ語2	タイ語3
1	図	ใช้ รูป ที่	พ ส่วน	ไว้ ใน รูป ที่
2	屋根材	วัสดุมุง	กระบวน การ ก่อ สร้าง ของ วัสดุมุง หลังคา	ไว้ และ วัสดุมุง หลังคา
3	端部	ปลาย ด้าน หนึ่ง	ไว้ ที่ ส่วน ปลาย	ส่วน ปลาย ยังสามารถ
4	錠剤	ยา เม็ด	ยา	เม็ดยา
5	内側ジョイント部	ส่วน ข้อ ต่อ ภาย ใน	ส่วนข้อ ต่อ ภาย ใน	แสดง ส่วนข้อ ต่อ ภาย
6	電力	กำลังไฟฟ้า	บัส กำลัง	ไฟฟ้า เป็น วัตต์-ชั่วโมง
7	入力電圧	การ รับ แรงดัน อินพุต	มาก กว่า แรงดัน อินพุต	และ แรงดัน อินพุต
8	クラッチ部材	ชิ้น ส่วน คลัตช์	ของ ชิ้น ส่วน คลัตช์	อยู่ บน ชิ้น ส่วน คลัตช์
9	ジョイント	ข้อ ต่อ	ข้อ	ในขณะที่ ข้อต่อ
10	電力トラック筐体ア	โครงสร้าง ราง สายไฟฟ้า	โครงสร้าง ราง สายไฟฟ้า แล้	โครงสร้าง ราง สายไฟฟ้า ยึด
11	駆動側クラッチ板	จาน คลัตช์ ด้าน ขับ เคลื่อน	ขับ เคลื่อน	
12	開口部	ส่วน ช่อง เปิด	ที่ มี ช่อง เปิด	เปิด ที่
13	蒸気	ไอที่	3 ไอที่	ไอน้ำ ได้รับ
14	実施形態	รูปแบบ การ ประดิษฐ์	รูปแบบ การ ประดิษฐ์ ที่	ใน รูปแบบ ต่างๆ และ
15	搬送体	ตัวหิ้ว	ของ ตัวห ิ้ว	ใน ของ ตัวห ิ้ว
16	出力部	เอาต์พุต ที่	เข้ากับเอาต์พุต	ไปยังเอาต์พุต ของ
17	屋根材本体	ตัว วัสดุมุง หลังคา	และ ตัว วัสดุมุง หลังคา	จะ รวม ถึง ตัว วัสดุมุง หลังคา
18	表面	ผิว	พื้นผิว	ได้รับ การ ปฏิบัติ ที่ พื้นผิว
19	特許文献	เอกสาร สิทธิบัตร	ดู สิทธิบัตรอ้าง อิง ที่	ใน สิทธิบัตรอ้าง อิง ที่
20	開口	รูเปิด	เคลื่อน ผ่าน ช่อง เปิด	ใน อุปกรณ์ ก่อน อื่น เปิด
21	フィルタ材	วัสดุ ไส้กรอง ที่	วัสดุ ไส้กรอง	ได้ วัสดุ ไส้กรอง ที่
22	電圧	แรงดัน	แรงดันไฟฟ้า	และ แรงดัน
23	筐体	ถ	ตัวราง มา	ไป ด้วย
24	電力変換器	ตัวแปลงผัน กำลัง	ตัวแปลงผันกำลัง	อุปกรณ์ ตัวแปลงผัน
25	箱体	รอบนอก ของ มัน	รอบนอก ของ มัน ที่	กล่อง
26	発明	ประดิษฐ์ ดังกล่าว	การ ประดิษฐ์ ดังกล่าว	ได้รับ การ ทำให้
27	被動側クラッチ板	คลัตช์ ด้าน	โดย ที่ จาน คลัตช์ ด้าน ถูก	คือ โดย ที่ จาน คลัตช์ ด้าน ถูก
28	缶エンド	ส่วน ปลาย ของ กระป๋อง ที่	ส่วน ปลาย ของ กระป๋อง	แผงกลาง ส่วน ปลาย
29	タブ部	ส่วนแถบ	ทำให้ ส่วนแถบ	ส่วนแถบ ที่
30	刃	ใบมีด	เปลี่ยน ใบมีด	ใส่ กับ ถอด ใบมีด

（4）英泰対訳辞書（一部）

#	英語	タイ語1	タイ語2	タイ語3
1	roofing material	หลังคา	เช่น วัสดุมุง หลังคา	หลังคา A
2	power	กำลัง	กำลัง เท่านั้น	ให้ กำลัง
3	part	โดย ที่ ส่วน	(tip	ง ประสงค์ ต่อไป ส่วน
4	housing	ตัวโครง	ตัวเรือน	ไว้ ล่วงหน้า
5	tablet	นี้ เป็น ลักษณะ ที่ ตัดยา	จ่าย ยา	แยก ส่วนยา เม็ด ที่ ผ่าน มา
6	roofing material body	ตัว วัสดุมุง หลังคา อีก ตัวหนึ่	ตัว วัสดุมุง หลังคา	หลังคา
7	opening	เปิด	ช่อง เปิด	ใน การ บำบัด ความ ผิดปกติ ทาง ประสาท วิทยา
8	water	ซับ น้ำ	ด ซับ น้ำ	ไหล ของ น้ำ มัน แบบผัน
9	input voltage	แรงดัน อินพุต ที่	แรงดัน อินพุต	ของ พอลิเมอร์
10	material	มี วัสดุ	ที่ มี วัสดุ	ไว้ ด้วย วัสดุ ที่ อยู่ ตัว ด้วย
11	raw material	วัตถุดิบ ไป	วัตถุดิบ ไป อยู่	ไป ได้ ที่ วัตถุดิบ
12	surface	แบน เรียบ พื้นผิว	ตามพื้นผิว	ทำให้ พื้นผิว
13	side	ด้าน ข้าง ที่	ด้าน ข้าง	แผง ข้าง
14	filter material	วัสดุ ไส้กรอง ตาม	วัสดุ ไส้กรอง ที่	วัสดุ ไส้กรอง เพื่อ
15	portion	ส่วน เคลื่อน ที่ เลื่อน ดังก	ส่วน เคลื่อน ที่	• ส่วน
16	inner joint part	ส่วนข้อ ต่อ ภาย ใน แบบ	ส่วนข้อ ต่อ ภาย ใน แบบ ดั้งเดิม	แสดง ส่วนข้อ ต่อ ภาย ใน ที่
17	end	ปลาย	ปลาย ด้าน	ไห้ หัน ส่วน ปลาย ด้าน
18	power track housing assembly	โครงสร้าง ราง สายไฟฟ้า	โครงสร้าง ราง สายไฟฟ้า ที่ ต่อ	โครงสร้าง ราง สายไฟฟ้า โดย
19	clutch member	คลัตช์	ชิ้น ส่วน คลัตช์	ที่ ชิ้น ส่วน คลัตช์
20	blade	ประคอง ใบมีด	ประคอง ใบมีด ดังกล่าว	ไป ข้าง หน้ากับ
21	embodiment	รูปแบบ การ ประดิษฐ์	รูปแบบ การ ประดิษฐ์ ที่	ไป ถึง ใน ตัวอย่าง การ ใช้งาน ซึ่ง
22	inner surface	พื้นผิว ภาย ใน	พื้นผิวด้าน	
23	control terminal	ขั้ว ต่อ ควบคุม ไว้	ขั้ว ต่อ ควบคุม ไว้ ที่	ขั้ว ต่อ ควบคุม
24	present invention	การ อธิบาย การ	อธิบาย การ	แผ่นฐาน ซึ่ง
25	voltage	แรงดันไฟฟ้า	แรงดัน	แรงดันไฟฟ้า ที่ ใช้ ใน
26	output terminal	ขั้ว ต่อเอาต์พุต	ต่อเอาต์พุต	
27	power converter	ตัวแปลงผัน กำลัง	ตัวแปลงผันกำลัง	
28	one	ตาม หนึ่ง	สุด หนึ่ง	ไป หน้าเจ็ด ค่า
29	fabric	ผ้า	ตึง ของผ้า	ไม่ ถัก ไม่ ทอชนิด
30	tablet splitting app	เม็ด เป้าหมาย การ แยก ส่	เม็ด เป้าหมาย การ แยก ส่วน ซึ่ง ถูก	และ