平成 24 年度

中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査

調 査 報 告 書

概 要 版

平成 25 年 3 月

特 許 庁

（空白）

目次

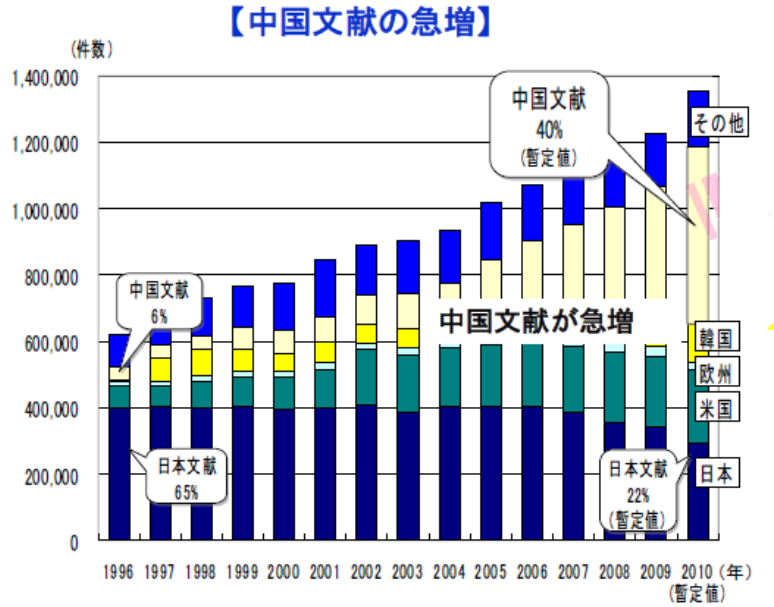
（空白）

1. 調査目的

従来から特許出願が多くアクセスの必要性が高かった英語などに加え、非欧米諸国の特許出願、特に中国からの出願が近年増大している。

英語文献については、従来から我が国審査官および出願人が原語でその発明の内容を理解することが可能であるが、近年重要性を増してきた中国語については、原文のみからの正確な理解は困難であると考えられる。また、中国の公開特許公報及び実用新案公報などの特許関連文献の年間発行数は既に米国特許公報、米国公開特許公報、欧州公開特許公報を上回っており、今後さらに増加するものと見込まれる。

図 1.1 (出典)産業構造審議会 平成 24 年 6 月 25 日配布資料



(出典)特許庁作成

そして、世界で通用する安定した権利を設定するためには、日本語、英語はもとより、それ以外の外国文献についても漏れなく調査することが必須となっている。出願件数の増加を背景に、中国での知財民事訴訟件数は米国を越え、なおも増加傾向にある。

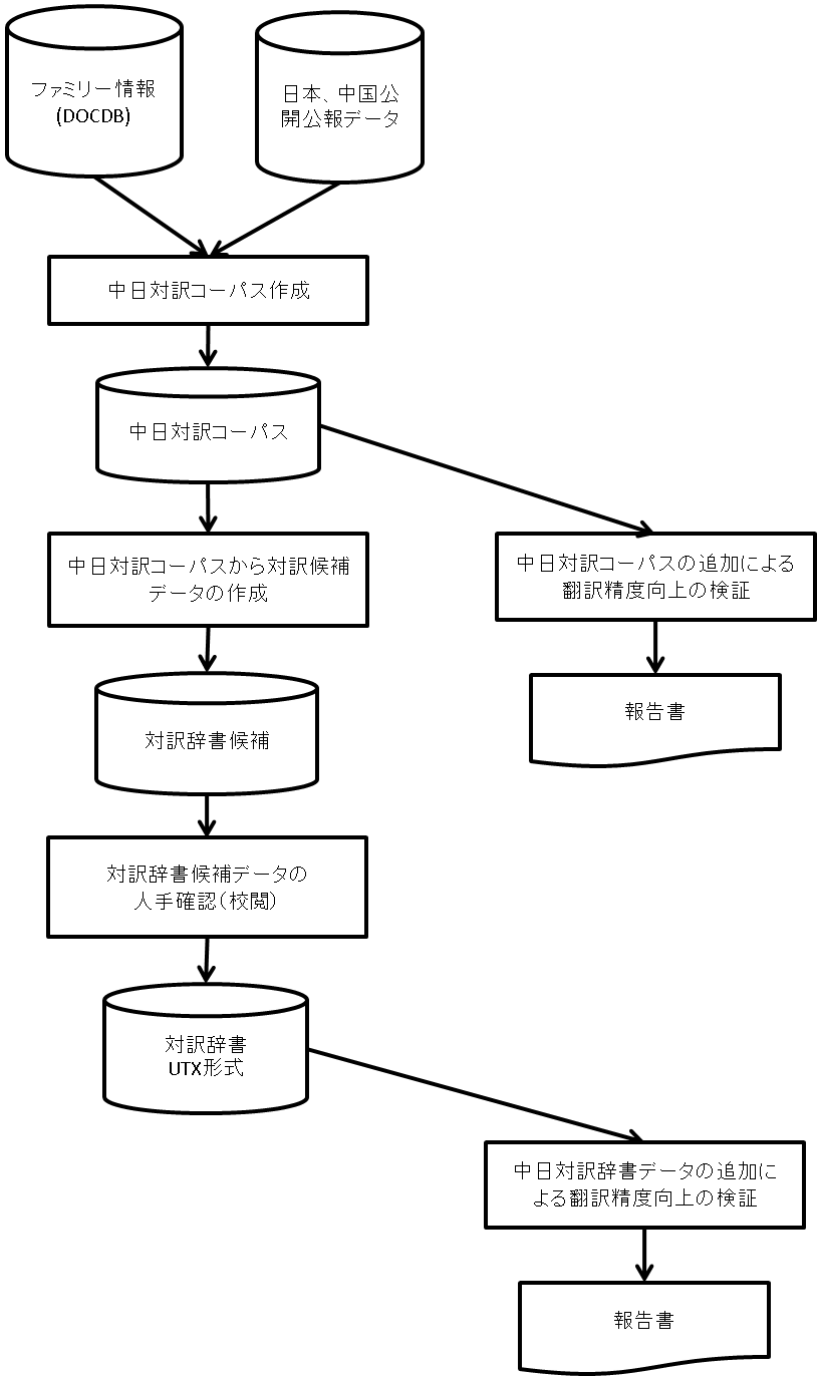
また、中国の特許関連文献が引用文献として引用される割合は他の特許庁において既に高く、技術的内容から見ても無視できないものとなっていることから、中国語の特許関連文献へのアクセス確保が急務となっている。このような状況を受け、中国語で記載された特許関連文献に対して機械翻訳を活用し、日本語により調査・理解可能とするインフラの早急な整備が必要不可欠となり特許庁では中韓文献翻訳・検索システムのリリースが今後予定されている。

このような中国の特許文献に対して日本語でアクセス確保するための言語資源として、特許用の中日辞書および中日対訳コーパスが有用である。

2. 調査概要

本調査では、平成 23 年度に特許庁が実施した「特許文献の機械翻訳のための辞書整備に関する調査」において、パテントファミリーからの効率的な辞書作成方法についての見解が得られたことを受けて、中国公開特許公報と日本公開特許公報のパテントファミリーから中日対訳コーパスを作成し、そこから統計的に対訳辞書候補データを作成し、専門家による用語の確認を経て中日対訳辞書データ 100 万語を作成した。

図 2.1 パテントファミリーからの辞書作成フロー



さらに前記調査で作成した中日対訳辞書データをルールベース機械翻訳システム(RBMT)に基づく市販の中日機械翻訳システムに適用した場合と、中日対訳コーパスを統計ベースの機械翻訳システム(SMT)に追加した場合の翻訳品質の向上について検証した。下記の3つに分けて調査の概要を説明する。

(2. 1)中日機械翻訳用辞書データの作成
(2. 2)中日対訳辞書データの追加による翻訳精度向上の検証
(2. 3)中日対訳コーパスの追加による翻訳精度向上の検証

なお、本調査は RBMT と SMT の翻訳精度の比較を目的としない。RBMT の検証は特許文書から作成した中日対訳辞書データを追加した後の精度向上を確認することを目的としている。その「中日対訳辞書データ」のデータソースには本検証の対象となった文献の半数(パテントファミリーがある案件)が含まれている。つまり、本検証は実際の使用時(つまり、辞書のデータソースとは無関係の文の翻訳)と比べて有利な条件で実施されている。そのため、RBMT の中日対訳辞書データ追加後の評価結果の数値は、実際よりも高い値となっている可能性がある。

SMT の検証は中日対訳コーパスの量を増やすと、どの程度翻訳精度が上がるかを確認することを目的としている。SMT は評価対象に含まれる文対を学習してしまうと、その文をそのまま出してしまう可能性があるため、評価対象文を除いて学習した。つまり、SMT の検証は実際の使用時と比べて有利な条件でない。また SMT では将来行う可能性がある調査との比較を分かりやすくするため、使用する中日対訳コーパスの量を 100 万件と 1,000 万文対とした。このコーパス数は対訳辞書作成で利用した 1,551 万文対より少ない。

このように RBMT と SMT の検証は条件が違うため、今回の調査結果の数値を比較して両者の優劣を判断することはできない。

2. 1 中日機械翻訳用辞書データの作成

中日対訳辞書データは以下に示す手順(a)～(f)で作成する。

(a)中日対訳コーパスの作成
(b)中日対訳コーパスからの対訳辞書候補データの作成
(c)対訳辞書候補データからの不要語の除去
(d)人手確認用対訳辞書候補データの人手確認(校閲)
(e)中日対訳辞書データの基本語と基本語以外の中日対訳辞書データの振り分け
(f)基本語の対訳辞書データ、基本語以外の対訳辞書データの UTX 形式への変換

手順(a)で中国と日本のパテントファミリーデータから中日対訳コーパスを作成し、(b)と(c)で対訳

辞書候補を機械的に抽出・不要語を除去し、それを(d)で専門家がチェック(校閲)することで短期間かつ低コストで中日対訳辞書データを作成した。ルールベース機械翻訳に適用するために(e)作成した辞書から基本語と基本語以外を分離した。作成した中日対訳辞書は(f)で汎用的な辞書形式に変換した。(e)と(f)の採用により、機械翻訳システムのユーザが特許用機械翻訳辞書を全分野共通または分野別に利用できるようにした。

以下に(a)～(f)のステップ概要を説明する。

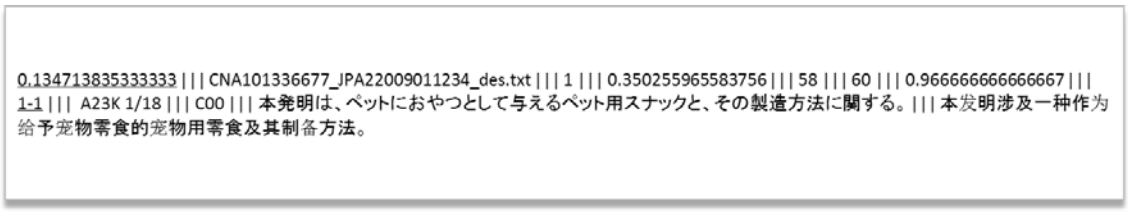
(a) 中日対訳コーパスの作成

DOCDB に蓄積されているパテントファミリー情報(family-id)を利用して、特許庁から貸与された2005年～2009年の中国公開特許公報(約105万件)と技術内容が対応する日本公開特許公報のリスト(約27万件)を作成し、それらパテントファミリーから中国語と日本語の文と文を対訳にした対訳文対の集まりを作成した。

中日対訳コーパスの作成には対訳関係にある中国語と日本語の文章を分析して、対応する文を自動的に決定する Japio が保有する中日自動文アラインメントツールを利用した。この中日自動文アラインメントツールが作成した中日文アラインメントを統計翻訳に用いる中日対訳コーパス及び中日辞書作成に利用することから、用途に合わせた文対応が選定できるように対応する文数(中国文と日本文の対応が1対0、1対1、1対n等)とそのツールが備えている中日対訳辞書に基づくアラインメントの対応度(文対応スコア)を付加した。

図 2.2 に作成した中日対訳コーパスの例を示す。1行目の先頭の数値が文対応スコア、2行目のの先頭の 1-1 が中国文と日本文が1文対1文で対応することを示す。

図 2.2 中日対訳コーパス(イメージ)



ツールの出力として得られた中日対訳コーパスは発明の名称・要約・請求項・明細書を合わせて 6,700 万文対である。表 2.1 に4分野の内訳を示す。

表 2.1 分野別中日対訳コーパス作成件数

分 野	件数
化 学	24,541,611
電 気	16,317,334
機 械	8,651,986
物 理	17,481,140
合 計	66,992,071

(b) 中日対訳コーパスからの対訳辞書候補データの作成

本処理は中日対訳コーパスから統計的に対訳辞書候補を抽出する処理である。処理は中日対訳コーパスから精度の高い中日対訳コーパスを抽出し、その日本語文と中国文をそれぞれ形態素解析して単語に分割した。その中国と日本の単語分割結果を統計的に学習して単語とフレーズの対応付けデータを作成した。さらに、この対応付けデータから日本語の専門用語を手掛かりに対応する中国語を抽出して対訳辞書候補データを作成した。

利用した中日対訳コーパスは(a)で作成した中日対訳コーパスから中国語文と日本語文が1文対1文で対応 かつ スコアが 0.1 以上の請求項と明細書から抽出した中日対訳コーパスである。スコア 0.1 以上としたのは Japio の日米対訳コーパスからの辞書抽出の経験に基づいた。名称は語数が少なく一部の単語の一致に影響されて誤った対応が排除できない可能性があるため処理対象から除外した。要約は日本と中国で内容が対応しない場合が多いため処理対象から除外した。表 2.2 に実際に対訳辞書候補データ作成に使用した4分野の中日対訳コーパスの件数を示す。

表 2.2 中日対訳コーパス絞込み後の件数

分 野	絞り込み後件数
化 学	5,320,986
電 気	4,219,635
機 械	2,182,762
物 理	3,792,471
合 計	15,515,854

(c) 対訳辞書候補データからの不要語の除去

(c)の工程は次の(d)校閲工程の効率を向上するための工程である。(b)対訳辞書候補データは統計的に処理したデータであるため、日本語として意味を成さない語や、対応する中国語が明らかに誤りである場合がある。例えば、日本語の語頭に「前記、上記、該」等の特許文書で良く表れる文字が余分につく場合、中国語の語頭に「在、的、了」等の名詞に含むべきでない部分が余分に付く場合等がある。(図 2.3 参照)これらの余分な部分を機械的に除去すること等の対応で、次の

人手確認工程の効率を向上することができた。

図 2.3 対訳辞書候補の中国語の誤り例

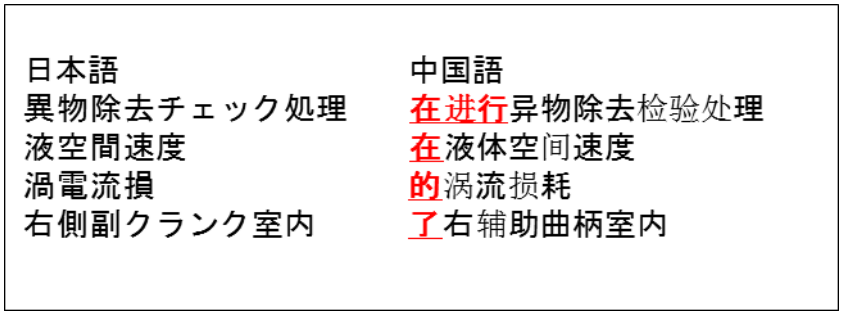


表 2.3 に機械処理の結果得られた4分野の対訳辞書候補数の内訳を示す。

表 2.3 対訳辞書候補数データ作成件数

分 野	対訳辞書候補数
化 学	1,582,416 語
電 気	1,229,012 語
機 械	764,671 語
物 理	1,208,814 語
合計(分野間重複なし)	4,460,048 語

(d) 人手確認用対訳辞書候補データの人手確認(校閲)

(c)の不要語の除去を行った後の対訳辞書候補データを人手により確認した(以下「校閲」という)。校閲作業は図 2.4 人手確認用対訳辞書候補データの中日対訳コーパス中の日本語と中国語の共起頻度の多い順で行った。この結果、約 145 万件の辞書候補データを校閲して 100 万語の中日対訳辞書データを得た。

図 2.4 人手確認用対訳辞書候補データ(イメージ)

	採用 不採用	品詞	日本語	中国語	日本語例文	中国語例文
1			アーチ状	拱状	例えば、Ｖ字状、**アーチ状**等である	例如成V字状、**拱状**等。

(e) 中日対訳辞書データの基本語と基本語以外の振り分け

(d)で得た 100 万語の中日対訳辞書データを市販の機械翻訳システムの基本辞書と中国語見出しでマッチングして基本辞書に含まれる基本語(25,077 語)と基本語以外の語(974,923 語)に分けた。

多くの場合、市販の機械翻訳システムの基本辞書の情報は後から追加する辞書に比べて充実しているので、基本語をユーザ辞書から除くことで翻訳の品質を向上させることが期待できる。言い換えると、基本語以外をユーザ辞書として使用することで、特許用の辞書を追加した場合の副作用が回避できる。

(f) 基本語の対訳辞書データ、基本語以外の対訳辞書データの UTX 形式への変換

作成した対訳辞書データはアジア太平洋機械翻訳協会(Asia-Pacific Association for Machine Translation 以下、AAMT)の辞書データ形式の規格 UTX を利用した。UTX は文字コードが UTF-8 のタブ区切りのテキストデータであり単純であるが、ファイルの先頭部分(ヘッダー)に #ではじまる2行の辞書の列の説明(名称)を定義できる。(図 2.5 参照)

今回の定義では UTX の必須項目である src(中国語)、tgt(日本語)、src:pos(中国語品詞)、tgt:pos(日本語品詞)に頻度情報をユーザ定義として追加した。追加したのは中日対訳コーパスにおける語の出現頻度である。出現頻度は src と tgt の語が単独で現れる頻度と src と tgt の語が同時に中日対訳コーパスに現れる頻度の3種類である。この頻度を中日対訳コーパス全体、と4つの分野別(c:化学、e:電気:、m:機械、p:物理)の合わせて5種類算出した。

その結果 3種類 × 5種類 = 15の頻度情報 を作成した。下の例はヘッダーの2行の辞書の列の説明と4語の対訳である。

(UTX の仕様は http://www.aamt.info/japanese/utx/ を参照のこと。)

図 2.5 中日対訳辞書データ(イメージ)

#UTX 1.11; zh/ja; 2013-03-01; JPO

#src	tgt	src:pos	tgt:pos	freq-all-src		freq-all-tgt		freq-all-src-tgt	
	freq-c-src	freq-c-tgt	freq-c-src-tgt		freq-e-src	freq-e-tgt	freq-e-src-tgt		
	freq-m-src	freq-m-tgt	freq-m-src-tgt		freq-p-src	freq-p-tgt	freq-p-src-tgt		
组合物	組成物	noun	noun	176846	180073	170343	147482	148526	142337
	9210	10114	8716	7295	7798	6845	12859	13635	12445
端部	端部	noun	noun	157989	190932	144120	38678	53250	34139
	41815	46058	38201	51570	61617	47738	25926	30007	24042
控制部	制御部	noun	noun	114633	133468	110810	15896	17003	15438
	40023	48781	38919	18672	20262	17887	40042	47422	38566
反应混合物		反応混合物		noun	noun	111709	108845	105387	109382
	106635	103290	536	504	482	499	487	476	1292
	1219	1139							

2．2 中日対訳辞書データの追加による翻訳精度向上の検証

本調査にて作成した「中日対訳辞書データ」を中日機械翻訳ソフトウェアに辞書として追加することによる翻訳精度の向上効果について検証した。検証用サンプルデータには中国公開公報の要約 160 件を利用し、これを、「中日対訳辞書データ」を辞書に追加する前後の環境で機械翻訳して、その出力結果を比較した。

(1)検証の環境

(a) 検証に用いたルールベースの中日機械翻訳ソフトウェア

検証には、市販のルールベース方式中日機械翻訳ソフトウェア『The 翻訳エンタープライズ V15』(東芝ソリューション株式会社製)を用い、本ソフトウェア単体での機械翻訳結果と、これに機械翻訳辞書として「中日対訳辞書データ」を追加した環境での機械翻訳結果とを対比する形式で実施した。

(b) サンプルデータ

検証用のサンプルデータには、中国語特許文献の要約 160 件を使用した。IPC 第 8 版の各セクション(A～H)より、年代や出願人の異なる案件を 20 件ずつ選択した。なお、各サンプルデータは、別途人手で中日翻訳を行い、検証の際のリファレンスとして用いた。

(c) 中日対訳辞書データ

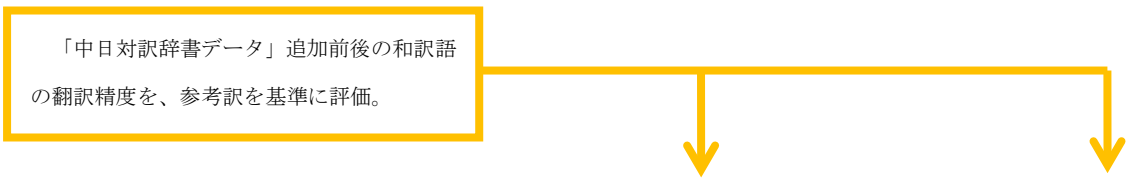
今回作成した「中日対訳辞書データ」は、各技術分野の特許文献から採取した中日対訳用語100 万語を収録しており、一つの中国語に対して複数の和訳語が収録されている場合がある。このため本検証では、「中日対訳辞書データ」を機械翻訳辞書として使用するにあたり、これを技術分野(化学、電気、機械、物理)ごとに 4 パターン作成し、一つの中国語が複数の和訳語を持つ場合、各技術分野にて最も使用頻度の高い和訳語のみを収録することとした。この処理により、「中日対訳辞書データ」から作成した機械翻訳辞書の収録語数は、4 分野とも 829,338 語となった。

(2)検証の手法

「中日対訳辞書データ」は名詞を中心に収録していることから、本検証は、サンプルデータ中で使用される全ての名詞を対象に、「中日対訳辞書データ」を機械翻訳辞書に追加する前後で、個々の名詞の訳語の精度が向上したか否かを評価した。具体的には、以下の手順にて検証を実施した。

1. サンプルデータに含まれる全ての中国語名詞をリストアップ。
2. サンプルデータの人手翻訳から各中国語名詞の和訳語をリストアップ(参考訳)。
3.「中日対訳辞書データ」追加前後の機械翻訳結果から実際の和訳語を抽出。
4. 参考訳に照らし「中日対訳辞書データ」追加前後の和訳語を評価。
5. 評価結果を集計し、各種評価値を算出。

図 2.6 参考訳に照らし「中日対訳辞書データ」追加前後の和訳語を評価



中国語	参考訳	「中日対訳辞書データ」追加前		「中日対訳辞書データ」追加後	
		和訳語	評価	和訳語	評価
本发明	本発明	本発明	○	本発明	○
据控制电路板	制御回路基板	制御回路板	○	制御回路板	○
单片机	ワンチップコンピュータ	シングル・チップ機	×	ディレクティブ	−
指令	指令	指令	○	−	×
移动	移動	シフト	×	移動	○
有人	人	人	○	人間	○
垃圾	ごみ	ジャンク	×	ゴミ	○
:	:	:	:	:	:

さらに補足評価として、サンプルデータ中の 40 件についてはデータ追加前後の機械翻訳結果全文を比較し、用語以外の部分の係り受け改善などの間接的効果の有無について調査した。

(3)検証結果

サンプルデータにあらかじめ含まれる全ての用語(名詞)について、機械翻訳ソフトウェア単体での翻訳結果(RBMT1)における和訳語と、そこに機械翻訳辞書として「中日対訳辞書データ」を追加した環境(RBMT2)における和訳語とをピックアップし、下記①～⑥の評価を行い、語数をカウントした。

① RBMT1 における各用語の訳語の正訳数
② RBMT2 において「中日対訳辞書データ」から訳語が採用された用語の数
③ 上記②のうち、RBMT1 から訳語が変化した用語の数
④ 上記③のうち、訳語が改善した用語の数
⑤ 上記③のうち、訳語が悪化した用語の数
⑥ RBMT2 における各用語の訳語の正訳数

下表に、サンプルデータ 160 件全件における、上記①～⑥の平均値を示す。

表 2.4「中日対訳辞書データ」翻訳精度検証結果平均値

	全用語数	①RBMT1 正訳	②RBMT2 採用	③RBMT2 訳語変化	④RBMT2 訳語改善	⑤RBMT2 訳語悪化	⑥RBMT2 正訳
語数	21.3 語	13.0 語	16.2 語	11.5 語	4.4 語	−0.8 語	16.5 語
%	−	61%	76%	54%	21%	−4%	78%

(注)本検証の対象となった用語は全て「中日対訳辞書データ」のデータソース中に含まれている語であり、通常よりも有利な条件で施行されている。このため RBMT2 に関する各評価結果の数値は実際よりも高い値となっている可能性がある。

上表のとおり、「中日対訳辞書データ」を機械翻訳辞書に追加する前の環境(RBMT1)における用語の正訳率は 61%であったのに対し、追加後の環境(RBMT2)ではこれが 78%に向上しており、このことから、「中日対訳辞書データ」を機械翻訳辞書に追加することにより、用語(名詞)の翻訳精度に関して一定の向上効果が得られることが確認できた。

また、本検証を実施した結果、以下のようなことがわかった。

・ 複合名詞ほど RBMT1 で誤訳となる割合が高く、ほぼ単名詞とみなせる中国語 2 文字以下の正訳率が 71%なのに対し、複合名詞とみなせる中国語 3 文字以上の用語の正訳率は 49%であった。RBMT2 ではこれがそれぞれ 81%、73%に改善しており、特に複合名詞の翻訳精度に大きな向上が見られた。

・ RBMT1 の時点で未知語はほとんど見当たらない。ただしこれは、実質的に未知語である語が 1 文字単位に分解されて翻訳された結果であり、意味的には完全な誤訳となる場合が多い。「中日対訳辞書データ」の適用で正しい単位に改善したケースも多く見られるが、完全には解消していない。

・ RBMT2 における翻訳精度の悪化は少ないが、「中日対訳辞書データ」から「過度に限定的な訳語」が採用されるケースや、多品詞を特定の品詞（名詞または動詞）で「中日対訳辞書データ」に登録した結果、RBMT1 では正しく解釈されていた多品詞が、RBMT2 では「中日対訳辞書データ」に登録した品詞に誤訳されるケースが見られた。

2．3 中日対訳コーパスの追加による翻訳精度向上の検証

本調査で作成した「中日対訳コーパス」を学習コーパスとして統計的機械翻訳システムに使用することによる翻訳精度の向上効果について検証した。検証は、中国公開特許の要約 160 件について、「中日対訳コーパス」中の 100 万文対を学習コーパスに使用した環境における機械翻訳結果と、これを 1,000 万文対に拡大させた環境での機械翻訳結果とを比較する形で実施した。

（1）検証の環境

(a) 検証に用いた統計翻訳ソフトウェア

本検証には、Japio と共同研究を行っている（独）情報通信研究機構（NICT）が開発した統計翻訳の一種であるフレーズベースの統計的機械翻訳ソフトウェア（詳細は添付資料 5.1 を参照）を使用し、本ソフトウェアを「中日対訳コーパス」の 100 万文対で学習させた環境における機械翻訳結果（SMT1）と、これを 1,000 万文対に拡大させた環境での機械翻訳結果（SMT2）とを対比する形式で実施した。

(b) サンプルデータ

本検証のサンプルデータには、2．2（1）に示した「中日対訳辞書データ」検証用のサンプルデータと同一のものを使用した。

(c) 中日対訳コーパス

SMT1 及び SMT2 にて中日統計翻訳ソフトウェアの学習に使用した「中日対訳コーパス」の文対は、以下の基準で選択した。

(1) 文対が1対1で、中国文の文字数が 400 文字未満の文対を全件抽出。
(2) 上記(1)より、スコア上位 2,500 万文対を抽出。
(3) 上記(2)より、ランダムに 1,000 万文対を抽出 ←SMT2 用学習コーパス

(4) 上記(3)より、ランダムに 100 万文対を抽出　←SMT1 用学習コーパス

※ ただし、サンプルデータに用いられた文献からの文対は（要約以外の箇所から得られた文対であっても）学習には使用していない。

(2)検証の手法

検証は、2．2(2)に示した「中日対訳辞書データ」検証の手法と同様の手法にて実施した。

(3)検証結果

サンプルデータにあらかじめ含まれる全ての用語（名詞）について、「中日対訳コーパス」100 万文対を学習コーパスに用いた環境における翻訳結果（SMT1）における和訳語と、学習コーパスを 1,000 万文対に拡大した環境（SMT2）における和訳語とをピックアップし、下記①～⑤の評価を行い、語数をカウントした。

① SMT1 における各用語の訳語の正訳数
② SMT2 において、SMT1 から訳語が変化した用語の数
③ 上記②のうち、訳語が改善した用語の数
④ 上記②のうち、訳語が悪化した用語の数
⑤ SMT2 における各用語の訳語の正訳数

下表に、上記①～⑤の 160 件全件の平均値を示す。

表 2.5「中日対訳コーパス」翻訳精度検証結果平均値

	全用語数	①SMT1 正訳	②SMT2 訳語変化	③SMT2 訳語改善	④SMT2 訳語悪化	⑤SMT2 正訳
語数	21.3 語	14.7 語	7.2 語	1.8 語	−0.7 語	15.9 語
割合	−	69%	34%	9%	−3%	75%

（注）本検証では、要約をサンプルデータに用いた文献からの文対は学習コーパスから除外している。このため「中日対訳辞書データ」の検証時とは条件が大きく異なり、両者を比較することはできない。

上表のとおり、「中日対訳コーパス」100 万文対で学習させた環境（SMT1）における用語の正訳率は 69%。コーパスを 1,000 万文対に拡大させた環境（SMT2）ではこれが 75%に向上した。上表①のとおり、学習コーパス拡大前の段階でも用語の翻訳精度は優秀であること、及びコーパスの拡大により一定の翻訳精度向上効果が得られることが確認できた。

また、本検証を実施した結果、以下のようなことがわかった。

- SMT1 で未知語扱いされた語の 23%が SMT2 で正しく訳されるようになった。
- SMT1 で誤訳された語（未知語を除く）の 31%が SMT2 で正訳となったが、その一方で SMT1

で正訳であった語が SMT2 で誤訳されるケースもあり、差し引きすると誤訳の改善率は 12%であった。

- 出力される翻訳文の文型（文構造）に関しては、SMT1 と SMT2 とで大きな変化は見られなかった。

2．4　事業実施スケジュール

調査スケジュールは平成24年8月10日に開始し、平成25年3月19日納品という約7か月間に100 万件の辞書作成とその効果の検証を行う必要があった。スケジュール的に最も心配された工程は対訳辞書の人手確認（校閲）であった。この前工程であるアラインメントの作成と辞書候補の抽出をできるだけ早く行ない、校閲工程に入り辞書の効果の評価の時間を確保することと校閲での用語の採用率を如何に向上するかが最大のポイントであった。

今回の調査では、Japio と（独）情報通信研究機構（NICT）が平成 22 年度から開始した中日機械翻訳の共同研究で得た知見と各種ツールを利用することで、校閲の前工程の機械処理の期間を短縮した。その結果、校閲に9月～1月の5か月間確保して 135 万件の校閲作業を行い、100 万語の中日対訳辞書データを作成した。その上で1月～2月の2か月間に作成した中日機械翻訳辞書データ及び中日対訳コーパスを利用した、翻訳精度向上の検証を行なった。

2．5 納品データ一覧

表 2.6 に本調査の結果作成された納品する電子データの一覧を示す。容量は gnuzip で圧縮後の容量である。

表 2.6 納品データ一覧

No	納品データ	型式	件数	容量 (byte)
1	基本語の中日対訳辞書データ (関連資料:添付資料 3.4)	UTX	25,080	2M
2	基本語以外の中日対訳辞書データ	UTX	974,924	76M
3	中日対訳コーパス			
3-1	中日対訳コーパス(発明の名称)	テキスト	267,703	16M
3-2	中日対訳コーパス(要約)	テキスト	723,525	163M
3-3	中日対訳コーパス(請求項)	テキスト	4,556,259	1.1G
3-4	中日対訳コーパス(明細書)	テキスト	61,444,584	12.0G
4	対応中国・日本公開特許公報番号リスト	テキスト	267,705	6M
5	対訳辞書候補データ	テキスト	3,029,917	115M