平成 23 年度

中国特許文献の和文抄録作成に対する機械翻訳の活用に関する調査

調 査 報 告 書 （概要版）

平成 23 年 11 月

特　許　庁

1. 本調査の目的

本調査は、中国語のみで公開されている中国公開特許公報を対象に実際に和文抄録作成事業を開始する際の課題や問題点を洗い出し、かつ翻訳精度を維持しつつ効率的に事業を実施する方策としての機械翻訳の利用可能性等について基礎調査を行うことを目的に行われた。

2. 調査の概要

本調査ではまず、和文抄録作成事業の対象と想定される、中国語のみで公開されている中国公開特許公報の件数の規模把握（件数調査）を行なった。さらに、これら中国語のみで公開されている中国公開特許公報からサンプル案件を選定し、公報全文から和文抄録を作成する方法、および対応する CPA（Chinese Patent Abstract:中国特許英文抄録）と出願人要約をそれぞれ人手翻訳および機械翻訳により和訳する方法にて、複数の中国和文抄録データのサンプルを作成し、それぞれの方法における抄録文としての品質について、翻訳精度、検索精度等の観点から評価・比較（品質評価）を行なった。

品質評価は、200 件のサンプル案件を選定し、それぞれについて 5 つの和文抄録作成方法にて 5 種類の中国和文抄録サンプルデータ（中国和文抄録）を作成したうえで、各和文抄録の品質評価を行い、その結果を分析・対比することで、各作成方法の有効性を相対的に評価した。さらに、調査の過程で発見された中国和文抄録に関する課題とこれに対する対処方法についての分析も行なった。

また、サンプル案件中 40 件について、4 種類の機械翻訳ソフトウェア／サービスを使用した 4 種類の全文機械翻訳サンプルデータを作成し、各種中国和文抄録サンプルデータとの品質比較を行うことで、機械翻訳を利用した公報全文翻訳データの実用性や課題についても分析した。

3.件数調査

中国和文抄録作成対象件数の規模把握のため、中国国家知識産権局（SIPO）へ出願され、2004～2008 年に公開された特許出願のうち、日米欧にて公開済みのパテントファミリーを持たない、すなわち中国のみで公開されている出願の件数を調査した。その結果を下表に示す。

＜中国出願のパテントファミリーの件数＞

発行年	SIPO発行件数	SIPO-JPO	SIPO-USPTO	SIPO-EPO	SIPOのみ
2004年	93,944	42,541	46,475	34,325	43,072
2005年	155,447	67,014	75,301	55,406	73,781
2006年	172,428	66,923	78,243	56,301	86,974
2007年	208,348	74,529	89,370	64,772	110,765
2008年	241,182	70,600	87,675	63,021	145,086

［技術分野別の件数分布］

中国のみで公開された出願については、さらに技術分野（IPC 第 8 版 A～H セクション）ごとの件数分布も調査した。以下にその結果を示す。

＜中国のみで出願公開された特許出願のセクション別件数＞

発行年	2004	2005	2006	2007	2008	合計
				5		5
Aセクション	10,397	18,602	17,430	21,287	23,775	91,491
Bセクション	5,632	9,536	9,605	12,397	17,625	54,795
Cセクション	7,718	13,579	16,375	20,621	28,013	86,306
Dセクション	968	1,537	1,945	2,186	2,996	9,632
Eセクション	1,364	2,549	3,611	4,984	6,167	18,675
Fセクション	2,917	4,878	6,245	8,328	10,461	32,829
Gセクション	7,613	12,483	13,850	17,796	23,963	75,705
Hセクション	6,463	10,617	17,913	23,161	32,086	90,240
合計	43,072	73,781	86,974	110,765	145,086	459,678

※　2007 年において IPC 分類が付与されていない文献が 5 件存在した。

4. 品質評価

【品質評価の対象】

品質評価は、2004 年～2008 年に中国のみで公開された特許出願から選定された 200 件のサンプル案件について、それぞれ下記 5 種類の中国和文抄録サンプルデータを作成し、これを品質評価の対象とした。

和抄#1 ： 中国語公開特許公報の全文読解による人手和文抄録作成
和抄#2 ： ＣＰＡ（英文要約）の人手翻訳
和抄#3 ： ＣＰＡ（英文要約）の機械翻訳
和抄#4 ： 中国語出願人要約の人手翻訳
和抄#5 ： 中国語出願人要約の機械翻訳

【品質評価の基準】

一次品質評価は、翻訳精度の観点、検索精度の観点、及び、和文抄録としての情報量の観点から、それぞれ 5 段階評価で採点する手法にて実施した。

【品質評価結果】

5 種類の中国和文抄録（和抄#1～#5）各 200 件の、翻訳精度・検索精度・情報量の各評価観点における品質評価結果（5 段階評価）の平均点と、その合計点は以下のとおり。

観　点	和抄#1	和抄#2	和抄#3	和抄#4	和抄#5
翻訳精度	4.9	4.6	3.1	4.8	2.4
検索精度	4.6	3.2	2.5	4.4	2.4
情報量	4.6	3.5	2.4	4.2	2.0
総合評価	14.1	11.3	8.0	13.4	6.8

5 種類の中国和文抄録中、最も品質が高いとされたのは現行の米国和文抄録と同等の方法で作成した和抄#1 であったが、中国語出願人要約を人手翻訳で和訳して作成した和抄#4 もこれとほぼ遜色のない高い評価を得た。

一方、CPA を人手翻訳した和抄#2 は、一部 CPA における品質面での問題（英語の品質、内容の省略や欠落）により、ほぼ同等の条件で作成された和抄#4 と比べると低い評価にとどまった。

機械翻訳を使用した和抄#3 と和抄#5 との比較では、CPA を英日機械翻訳した前者が、中国語出願人要約を中日機械翻訳した後者よりも優れていると評価された。それぞれを人手翻訳した和抄#2 と和抄#4 との比較では、後者が前者よりも高く評価されていたことを考えると、この逆転現象は、現状の中日機械翻訳の性能の未熟さを示しているといえる。なお和抄#3 及び和抄#5 は、今回の評価結果は上限値ではなく、機械翻訳技術の進歩により徐々

に向上していくことが予測される。また、ソース文献の言語品質や情報量、そして中日機械翻訳技術の今後の伸び代を考えると、将来的には中日機械翻訳を利用した和抄#5 のほうが高品質となる可能性が高い。

【機械翻訳の補助的な活用の可能性】

品質のみを見た場合、機械翻訳のみを使用して作成された和文抄録は、人手翻訳を大きく下回るという評価となった。しかしながら、機械翻訳を人手翻訳と組み合わせ補助的に利用することで、和文抄録の品質の向上や作成作業の効率化が果たせる可能性はある。

現状で特に有効な活用案としては、①和文抄録の補足情報としての機械翻訳結果の提供、及び、②人手翻訳の支援情報としての機械翻訳の活用、が推奨される。

[現状の課題と対処方法]

詳細品質評価では、中国和文抄録の作成に関する現状の課題として、以下の 5 つの課題を挙げ、その内容の分析と対処方法の検討を行なった。

課題 1：中日機械翻訳の現時点での性能面での不足

→　今後、技術用語辞書が充実し、さらに中日機械翻訳エンジンに英日機械翻訳と同程度の翻訳規則が実装されれば、中日機械翻訳を使用して実用域の和文抄録が作成できる可能性は充分にある。現状においてユーザ自らが実行できる対処策として、技術用語辞書の整備がきわめて有効と考えられる。

課題 2：CPA の品質

→　CPA は中国語の出願人要約の英訳版であるが、実際に両者の内容を比較してみると、CPA にはかなりの情報の省略や欠落が生じていることが判明した。本調査の対象案件 200 件全件の和抄#2（CPA の和訳）と和抄#4（中国語出願人要約の和訳）の平均文字数を比較した結果、和抄#2 が約 230 文字、和抄#4 が約 367 文字と、顕著な差異があった。

課題 3：翻訳者に日本語ネイティブ以外を起用した場合の品質維持

→　中国和文抄録作成に日本語ネイティブでない翻訳者を起用する場合、日本語の精度が低下する懸念がある。不適切な技術用語や誤記を洗い出してアラートする機械的な校閲が効率的である。

課題 4：「実施例」に関する情報の取得

→　現行の米国和文抄録は「実施例」に関する記載を要件としているが、中国語出願人要約や CPA はこれを記載要件としておらず、これらをソース文献として中国和文抄録を作成した場合、「実施例」に関する情報が欠落する懸念があった。しかしながら、本調査にて各種中国和文抄録の「実施例」に相当する情報の多寡を評価した結果、少なくとも中国語出願人要約については、ある程度の情報は含まれている場合が多いと判明した。

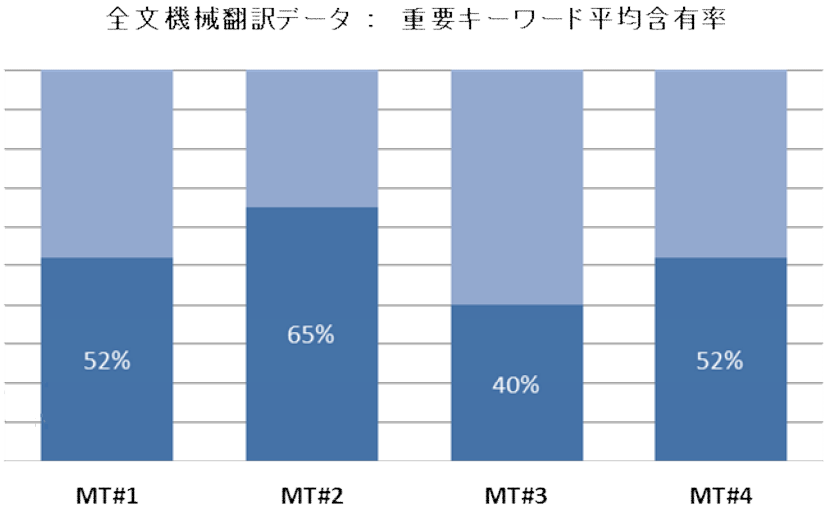
課題 5：データ処理における課題と留意点

→　中国特許公報データには、全文のイメージデータである TIFF イメージデータと、XML テキストデータとがあるが、後者において化学式などの情報の欠落が若干数検出された。また後者は改行タグ等の文中タグを含むため、機械処理の際は適切な前処理が必要となる場合がある。

［中国和文抄録データと全文機械翻訳データの品質比較］

追加調査の一環として、一次調査のサンプル案件 40 件それぞれについて、4 種類の異なる中日機械翻訳ソフトウェア／サービスを用いて作成した下記 4 種類の公報全文機械翻訳サンプルデータと、一次調査で作成した中国和文抄録の各種サンプルデータとの品質比較を行い、機械翻訳を利用した公報全文機械翻訳データの有効性や課題について分析した。

その結果、特に検索精度の観点で、統計翻訳方式による全文機械翻訳データ（MT#2）がルールベース方式による他の全文機械翻訳データを上回る結果となった。検索精度の評価は、重要キーワードが適切な訳語で含まれているか否かで判定するが、下に示すとおり、その含有率は MT#2 が特に高い。



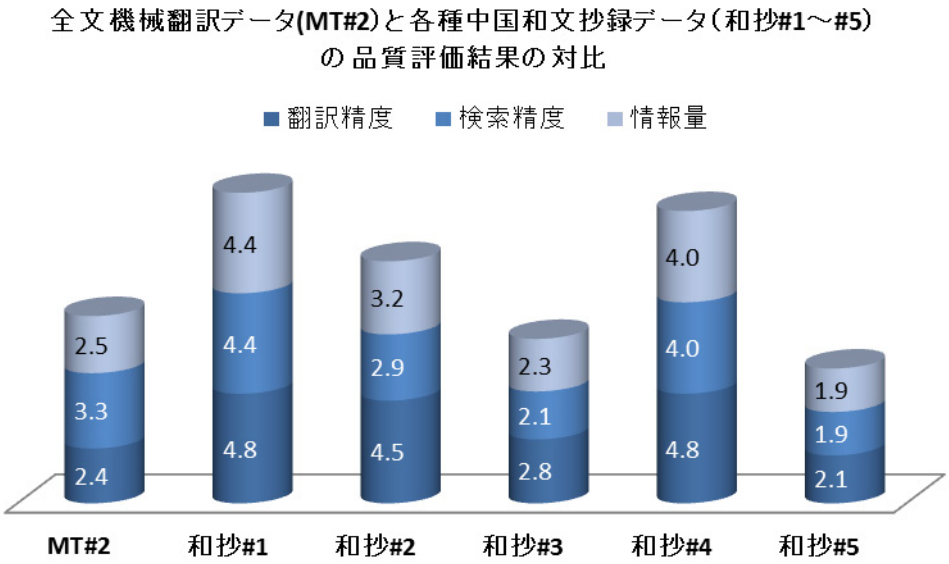
この結果は、裏を返せば、現状の中日機械翻訳において技術用語辞書の整備が進んでいないことを示すものといえる。MT#2 以外の 3 種の全文機械翻訳データは、いずれもルールベース方式の機械翻訳エンジンを用いているが、ルールベース方式の場合、技術用語辞書の貧弱さは技術用語の訳質に直結する。これに対し、MT#2 は統計翻訳方式の機械翻訳エンジンにより作成されている。統計翻訳方式は、ある程度の量の多言語コーパスが確保できれば直ちに性能を発揮できるため、発展途上段階にある現状の中日機械翻訳においては、技術用語の訳質という面では特に有利であるといえる。

ただし、今回の結果はあくまで現時点でのものであり、実務での使用を考える以上、そ

れぞれの本質を踏まえた検討が必要である。統計翻訳方式とルールベース翻訳方式との比較は日英／英日機械翻訳でも繰り返し行われてきたが、両者の本質的な傾向として、個々の用語の自然さでは統計翻訳方式が上位だが、文法的な正確さはルールベース方式のほうが上、という評価が一般的である。今後、中日機械翻訳が英日／日英機械翻訳と同様に発展を遂げる過程では、こうした統計翻訳方式、ルールベース方式それぞれの本質的な特徴が徐々に顕在化してくるものと予測される。

【全文機械翻訳データと中国和文抄録データの品質比較】

全文機械翻訳データで最も高品質とされた MT#2 と、各種中国和文抄録データ（和抄#1～#5）の品質評価値の比較結果を以下に示す。



上表のとおり、全文機械翻訳データは、人手翻訳を利用した和抄#1、#2、#4 には総合的な品質において及ばないものの、同じく機械翻訳を用いて要約情報を提供する和抄#3 や和抄#5 よりもその実用性は高く、また、こと検索精度の観点においては、CPA を人手翻訳した和抄#2 よりも高品質との評価を得た。MT#2 自身の評価値も検索精度に関する部分が最も高く、全文機械翻訳データは、特に検索用途に適しているといえる。

なお、和抄#1 や和抄#4 の検索精度の評価点は MT#2 の評価点よりも大幅に高いが、このことは必ずしも、和抄#1 や和抄#4 を提供すれば MT#2 から得られる検索上のメリットは全てカバーされるということを意味しない。全文機械翻訳データには、ソース文献が全ての選定キーワードを含んでいるという強みがあり、実際、サンプル文献においても、和抄#1 や#4 で欠落していた技術用語が MT#2 でカバーされているというケースも見られた。従って、和抄#1 や和抄#4 を採用する場合でも、検索用途に優れた全文機械翻訳データを補助的に提供することには大いに意味がある。