中国公開特許公報の日本語への機械翻訳

普及支援課特許情報企画室調査班調査第二係　　船守　茉美

抄録

アジア各国における出願件数の急伸から、その特許情報にも注目が集まっています。なかでも、出願件数が急伸している中国の文献は特に、ニーズが高まっています。しかしながら、中国の文献を原文のまま解することは、審査官にとっても出願人にとっても労力を必要とするでしょう。そこで、中国語で記載された特許文献を日本語で理解するために、機械翻訳の活用が期待されています。本稿では、平成21年度に実施した「中国公開特許公報の機械翻訳による日本語での提供に関する調査」を中心に、中国語を対象とした機械翻訳の現状をご紹介します。

1. はじめに

近年アジアの驚異的な経済的躍進に伴い、アジア各国における特許出願件数が急伸し、その特許情報にも注目が集まっています。特に、中国の特許出願件数は2009年には約31万5千件となり、2004年比で約2.4倍以上にもなっており、その出願件数の急伸ぶりがうかがえます[1)]。また、中国の公開特許公報及び実用新案公報等の特許関連文献が引用文献として引用される率は他庁において既に高く、技術的優位性を有するにもかかわらず、中国語でしか公開されない特許関連文献の割合が増加していることも指摘されていることから、我が国審査官および国内産業界においてニーズが高いものとしてアクセス性の確保を急いでいます。

このように日本国特許庁（以下、「JPO」という。）において、網羅的な先行技術調査と安定性の高い権利付与を維持していくためにも、量（出願件数）という観点からだけではなく、質（技術的優位性）の観点からも、JPOの審査官が中国特許文献に容易にアクセスできるための手段の確保が求められています。

しかしながら、中国語で記載された文献を原文のまま解することは、審査官にとっても非常に困難でしょう。また、我が国出願人にとっても英語文献と比して、侵害調査などに必要な読解のための労力は大きなものです。このように、出願人や我が国審査官が中国語等の非英語で記載された特許文献を扱う機会は増加しているにもかかわらず、先行技術文献調査や侵害調査等において言語の違いは高い壁となって厳然として立ちはだかっています。このような状況を受けて、すべての審査官および出願人が外国語で記載された文献に対して、日本語で容易にアクセスするための手段として、機械翻訳の活用が期待されています。

そこで、まず中国特許文献を巡る現状についてご説明し、中国公開特許公報の機械翻訳の整備に向けて、実際の中国公開特許公報テキストデータを用いて、記述方式・内容を分析し、特性の有無を洗い出すとともに、同データを機械翻訳するにあたって想定される問題点の把握、対応策の提案を目的に、平成21年度に実施した「中国公開特許公報の機械翻訳による日本語での提供に関する調査」の内容を中心にご紹介します。

2. 中国特許文献を巡る現状

中国特許文献のアクセス性確保が急務であることを物語る事件として、外国企業が中国国内の特許権の侵害で提訴され、敗訴したモトローラ、シュナイダー事件があります[2)]。この事件は、中国において、2005年にモトローラが、2006年には仏シュナイダー系列の企業がそれぞれ特許侵害で提訴され、何れも敗訴しているものです。モトローラは、中国の一市民が既に取得していた「回転式ディスプレイ携帯電話」の特許に対して、モトローラ製品が特許権を侵害していると提訴され、34万元の賠償額で敗訴しました。

また、シュナイダーは、中国の大手電機メーカが既に取得していた「高分断小型断路器」の実用新案に対して、既に販売済みのシュナイダー製品が侵害していると提訴され、3.3億元という中国の知財権侵害提訴では過去最大額で敗訴しました。

これら2点の訴訟が特許業界に与える影響は大きく、欧州の業界団体からEPOに対して、中国文献のアクセス性向上を要請するまでに至っており、ローカル文献（国際出

1) http://www.wipo.int/ipstats/en/statistics/patents/
2) 平成20年度特許庁検索システム最適化調査報告書　多言語横断検索技術に関する次世代検索システム開発に向けた調査　調査報告書　1-6～1-7頁

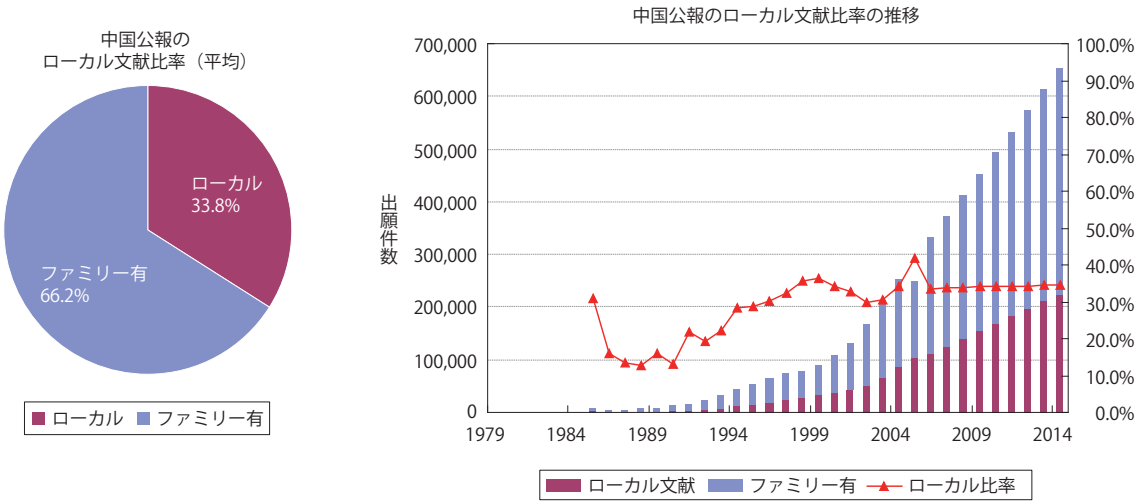


図1　中国公報のローカル文献比率の推移
出典：中国、韓国の予測値（特許庁作成資料）

願番号や優先権主張番号が存在しない文献）へのアクセス性を高める事の重要性は世界的にも高まっています。

上記の通り、ローカル文献へのアクセス性向上は重要事項ですが、中国のローカル文献比率は図1の推移を辿っており、平均すると中国では33.8%の割合となっています。これらのローカル文献は、英文抄録以上の内容を把握することが現状困難であり、約3割のローカル文献へのアクセス性向上は必須事項であると考えられます。

3. 調査内容

上記のような状況を受け、中国公開特許公報の機械翻訳の整備に向けて調査では、実際の中国公開特許公報テキストデータを用いて、公開特許公報の部分別（タイトル、アブストラクト、クレーム、ディスクリプション）、技術分野セクション（IPC第8版A～Hセクション）別に、一般的特徴および、定型化可能な表現の分析を行いました。また、分野別特性の有無、対応する日本語文献が存在する文献（以降、「通常文献」という。）とローカル文献の特性の有無について調査・分析を行うと共に、同データを機械翻訳するにあたって想定される問題点（異表記、表記のゆれ等）の把握、対応策の提案を行いました。

名詞句を正しくとらえる事が重要との考え方に基づき、特に、外来語などの未知語、組織名の異表記・表記のゆれ、数式・化学式・塩基配列について分析を行いました。

4. 分析ツールの紹介－自動アラインメント[3)]

中国特許文献と、対応する日本特許文献が与えられた場合、中国特許文献のテキスト一行が日本特許文献のどのテキスト一行と対応するかを自動的に判別することを目的とします。自動アラインメントの結果、対応が確認できた文は、目視によるチェックを経て、定型文として利用することが可能となります。また、結果をさらに別のツールで処理するなどして、単語単位やフレーズ単位での対応がとれれば、定型フレーズや用語の抽出も可能となります。

中国文と日本文が対応しているかどうかを判別する場合、同じ意味を持つ特徴的な表現が双方に含まれているかどうかを検査することが有力な手段となります。このような特徴的な表現を特徴要素と呼びます。特徴要素として一般的には数字列が有効とされており、本調査でも数字列を特徴要素として採用しました。

特許文献においては、参照番号の数字列が有効な特徴要素となっています。特許文献では図を用いて発明の詳細を説明することが多いですが、その場合、図の各要素を参照するために参照番号を用いており、参照番号は中国特許文献でも日本特許文献でも同一となり、対応する文で共起するためです。

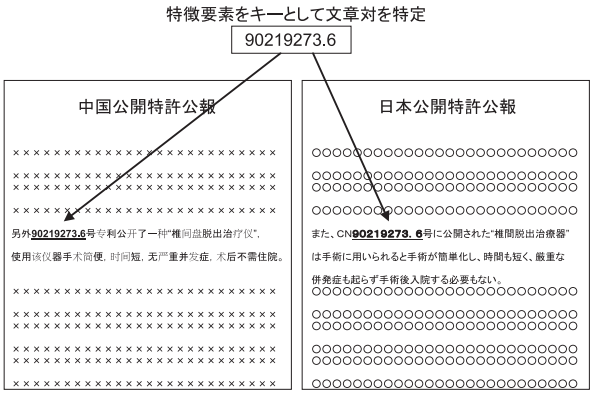
以下に、実際に自動アラインメントを行った結果を例示します。

■例

特徴要素	90219273.6
中国特許	另外90219273.6号专利公开了一种“椎间盘脱出治疗仪”，使用该仪器手术简便，时间短，无严重并发症，术后不需住院。
日本特許	また、CN90219273.6号に公開された“椎間脱出治療器”は手術に用いられると手術が簡単化し、時間も短く、厳重な併発症も起らず手術後入院する必要もない。

この例では、特徴要素は「90219273.6」となり、自動アラインメントは成功しています。これを図示したものを以下に示します。

3) 別の言語で表現された同じ内容の2つの文献が与えられた場合、文の対応をつけること。



5. 定型化可能な表現の分析

中国公開特許公報の各部分において、定型的に用いられる表現・フレーズを収集し、出現頻度の高い表現について分析を行い、定型化が可能な表現については対訳を付与します。定型文、定型パターン、定型フレーズを対訳で登録することにより、機械翻訳結果の品質向上が期待できます。

定型文は、複数の文献中に出現する文を指します。機械翻訳システムにおいては、翻訳メモリと呼ばれる機能を使用し、原文と訳文のペアを登録して、完全に一致する文が出現した場合に訳文を再利用することで、翻訳結果の品質向上を行うことができます。

定型パターンは、頻度の高いパターン文で、パターン文とは文の一部が変数となる文を指します。例えば、飲食店で良く使われる言い回しに「本日のメインディッシュは～です。」という文があった場合、「～」が変数ということになります。なお、変数部分以外を固定部分と呼びます。機械翻訳システムにおいては、翻訳メモリ機能を使用し、その原文のパターンと訳文のパターンを登録して、パターンに一致する文が出現した場合には、その訳文パターンを再利用し、変数部分だけを機械翻訳することで、翻訳結果の品質向上を行うことができます。

定型フレーズは、頻度の高いフレーズ（句）を指します。機械翻訳システムにおいては、名詞句・形容詞句・副詞句・動詞句など、単語より広い単位の句として辞書に登録することにより、機械翻訳結果の品質向上が期待できます。

5-1. 公開特許公報の部分別の分析

タイトル・アブストラクト・クレーム・ディスクリプションの4つのフィールドそれぞれについて4つのフィールドごとに特徴的な定型表現は、複数のセクション・通常文献・ローカル文献いずれにも使用される例が主であり、セクション固有の定型表現というのは、そのセクションに特有の語を含む例がほとんどでした。いずれも、高頻度の文に良質な訳文をつけて機械翻訳システムに登録することで、良質な訳文を再利用することができ、訳質向上につながると考えられます。

6. 技術分野別特性の分析

公開特許公報の本文に数式・化学式・塩基配列等が含まれる場合、この範囲をひとくくりに捉えられず、分断して構文を解釈することにより誤訳になることがあります。以下に例を示します。

原文	首先按Ｔ＝（Ｍ×Ｃ〈 sub 〉1〈 /sub 〉）/（Ｃ〈 sub 〉2〈 /sub 〉×ａ）计算所需的富硒红米的重量。
機械翻訳結果	真っ先にＴｈ＝（Ｍ×Ｃ〈sub〉１〈/sub〉）を押して/（Ｃ〈sub〉２〈/sub〉ｘａ）は計算するしなければならない裕福なセレンの紅の米の重量。

この例では、ひとくくりに捉えたい数式が、"/"で分断されてしまっています。

これらを翻訳対象外とすることで翻訳結果の品質向上が期待できます。そこで、数式・化学式・塩基配列の開始、中間、終端のパターンを分析して範囲の特定方法について検討を行いました。また、分野別の数式・化学式・塩基配列・有機化合物の出現割合を調査した結果、数式・化学式・塩基配列・有機化合物すべてにおいて、Ｃセクションの割合が非常に高く、このような範囲の特定が、特にＣセクションの文献の機械翻訳結果の品質向上に貢献できるであろうことがわかりました。

7. 異表記・表記のゆれの分析

機械翻訳においては、異表記・表記のゆれがある場合、既存の機械翻訳辞書ではそれらのバリエーションが網羅されていないため未知語または誤訳になる可能性が高くなります。特に組織名において異表記・表記ゆれが多いことが考えられるため、組織名を対象に異表記・表記ゆれについて分析を行い対応方法の検討を行いました。

次に外来語一般の表記のゆれの問題について分析と検討を行い、表記のゆれを吸収するため音が近く、入れ替わりやすい漢字セットからなる表である表記ゆれテーブルを作成しました。

7-1. 外来語の中国語表現と表記のゆれ

ここでは「意訳[4]」と「音訳[5]」の概念について説明しま

4) 漢字の意味の組み合わせで目的の単語の意味を表現する方法
5) 漢字の音（読み）を使って原語の発音を表現する方法

す。「意訳」と「音訳」は外来語の中国語表現のための主要な方法です。

(1) 意訳語

「インフレ」を「通货膨胀」とするように漢字の意味の組み合わせで目的の単語の意味を表現する方法です。語義文を短縮したものとも言えます。

固有名詞以外はかなりの場合、意訳します。例を表-1に示します。

表-1　意訳語の例

日本語	中国語
インフレ	通货膨胀
スキャナー	扫描器

特徴的なのは企業名でも意訳を使うことがあることです。例を表-2に示します。

表-2　企業名の意訳語の例

日本語	中国語	備考
サンマイクロシステムズ	太阳微系统	「サン」→「太阳」(太陽)、「マイクロ」→「微」、「システムズ」→「系统」。
フォルクスワーゲン	大众汽车	「フォルクス」→「大众」(大衆)、「ワーゲン」→「汽车」(自動車)。

(2) 音訳語

「盘尼西林」(ペニシリン) のように漢字の音 (読み) を使って原語の発音を表現する方法です。いわゆる「当て字」です。人名などの固有名詞では大抵は音訳が行われます。例を表-3に示します。

表-3　音訳文字を使っている音訳語の例

日本語	中国語	備考
イブプロフェン	布洛芬	薬品名
ケブラー	凯芙拉	商標
ルクス	勒克斯	単位

(3) 音訳／意訳複合

人名・地名にちなむ物の名前では人名・地名だけを音訳することがあります。これは日本語でもしばしば見られますが、日本語よりも意訳の度合いが高いです。音訳／意訳複合の例を表-4に示します。

表-4　音訳／意訳複合の例

日本語	中国語	備考
ベジェ曲線	贝塞尔曲线	「ベジェ」→「贝塞尔」(音訳)、「曲線」→「曲线」(意訳)
カルマンフィルター	卡尔曼滤波	「カルマン」→「卡尔曼」(音訳)、「フィルター」→「滤波」(意訳)

(4) 異表記

異表記とは同一の概念に対して複数の表記があるものを指します。新語の場合、まだ訳語が定まっていないことも多く、異表記が併用されます。これは時間とともに解消していく傾向にありますが、異表記が解消しないまま複数の表記が共存し続けることもあります。新語の異表記の例を表-5に示します。

表-5　新語の異表記の例

日本語	中国語
ブログ	网志 (意訳) ／博客 (中国大陸、音訳) ／部落格 (台湾、音訳)
アスパルテーム	阿斯帕坦／阿斯巴甜／阿斯巴坦

また新語でなくても、地域によって (特に中国大陸と台湾、香港などで) 異表記が定着していることもあります。ここでいう異表記とは簡体字／繁体字の違いではなく、中国大陸と台湾で全く別の表記になることをいいます。地域による異表記の例を表-6に示します。

表-6　地域による異表記の例

日本語	中国語
プリンタ	打印机 (中国大陸) 印表機 (台湾)

(5) 表記のゆれ

表記のゆれとは異表記のうち特に相互の差異が少ないものを指します。ここでは音訳語での漢字選択の違いによって発生する表記のゆれを主に取り扱います。簡体字・繁体字の違いによる表記のゆれもありますが、これは文字の対応が明確に決まっており機械的な変換が可能であるため、機械翻訳上は問題とならないと考えられます。

7-2. 組織名における異表記・表記のゆれ

通常文献の発明の出願人欄より組織名・人名の抽出を行った結果、最も異表記が多かった組織名では29の異表記が見られました。この異表記の多さは、組織の方針として、外国に出願する場合にはグループ会社は母体となる社号を利用するというルールがあるものと予想されます。

7-3. 異表記・表記のゆれの分析

異表記・表記のゆれの原因と考えられる事項を以下に示します。

(1) 法人格のゆれ

株式会社に対応する表記が、以下の4種類存在することが、表記ゆれの原因と考えられます。

公司
有限公司
股份有限公司
株式会社

法人格が社名の前後に付くことによるゆれもあります。これにより、同じ会社でもいくつかの表記が存在します。

表-7　法人格のゆれ

日本語	中国語
オムロン株式会社	欧姆龙株式会社
	欧姆龙公司
パイオニア株式会社	先锋电子株式会社
	先锋电子股份有限公司
	先锋公司
	先锋株式会社

(2) アルファベット表記・漢字表記による異表記

日本語カタカナ表記に対して、アルファベット表記・漢字表記の両方があります。

表-8　アルファベット表記・漢字表記による異表記

日本語	中国語
イビデン株式会社	IBIDEN 股份有限公司
	揖斐电株式会社
ジェイエスアール株式会社	JSR 株式会社
	捷时雅株式会社

(3) 社名の詳細表記による異表記

日本語または、または中国語の社名を詳細に表記している例です。

表-9　社名の詳細表記による異表記（日本語ベース）

日本語	中国語
ホシデン株式会社	星电株式会社
	星电器制造株式会社
株式会社日立製作所	日立空调系统株式会社
	株式会社日立制作所

(4) 単純な表記のゆれ

長音の表記のゆれ等による表記のゆれ。

(5) 社名変更等による異表記

社名変更等が行われたと思われる異表記。

(6) 音訳による異表記

複数の音訳が存在することによる異表記

(7) 意訳、音訳による異表記

意訳と音訳が存在することによる表記のゆれの例があります。具体的には、以下の例があります。

表-10　意訳、音訳による異表記

日本語	中国語
株式会社ミツバ	株式会社三叶 ※意訳　三叶草＝クローバー
	株式会社美姿把 ※音訳
スズキ株式会社	铃木发动机株式会社 ※发动机＝エンジン
	铃木汽车株式会社 ※汽车＝自動車
	铃木株式会社
富士写真フイルム株式会社	富士写真菲林株式会社 ※意訳　菲林＝フィルム
	富士写真胶片株式会社 ※意訳　胶片＝フィルム
	富士胶片公司
	富士摄影胶片株式会社 ※意訳　摄影＝写真を撮る
	富士胶片株式会社

(8) 複数訳語による異表記

1つの日本語に対応する中国語が、複数あることが原因による異表記です。以下の例があります。

日本語	中国語
浜松	滨松、浜松
通信	通讯、通信
並木	并木、並木
大塚	大冢 、大塚
紡績	纺织、纺绩
応化	应化、応化
鋼鈑	钢板、钢钣
コンピュータ	计算机、电脑

(9) 異なる出願人

調査の結果、中国語に対応する日本語が、全く異なる組織の場合もありました。

1つの中国語の組織名に対して複数の組織名が存在する場合、単純に異表記・表記のゆれのみで片付けることは困難です。

機械翻訳に利用する組織名辞書を作成する場合に一番の問題となるのは、対訳になっていない「(9) 異なる出願人」のケースでしょう。経過情報や単純な番号の間違えを考慮する機械処理は困難であり、最終的には人手による確認が必要となります。

7-4. 組織名辞書の作成

機械翻訳では異表記・表記のゆれがある場合、未知語あ

るいは誤訳の原因となるため、組織名の異表記・表記のゆれに対応する方法として考えられるのは組織名辞書の作成です。

本調査では、3,367件のデータから、1対1で対応しているものは1組ずつ、中国語ベースで日本語の異表記・表記のゆれがあるもの、日本語ベースで中国語の異表記・表記のゆれがあるものをすべて登録しました。

7-5. 外来語一般の表記のゆれへの対応

表記のゆれは組織名に限らず、音訳した外来語で一般に見られる現象です。組織名と違い分野は限定されないため網羅的な辞書を作ることは困難であり、別の手法が必要とされます。

そこで、表記のゆれへの機械翻訳の対応方法としては、あらかじめ「得」と「徳」のように音が近く、入れ替わりやすい漢字セットからなる表（これを「表記ゆれテーブル」と呼ぶ）を用意しておき、機械翻訳での辞書検索の際に、これを利用した曖昧検索を行うという方法が考えられます。本調査では、実際の文献から表記のゆれを収集し、作成を行いました。

8. 未知語の分析

未知語とは、一般に辞書に記載がない単語のことを指します。未知語は訳質悪化の原因となるので、辞書登録作業を継続して行うことで、その未知語の出現頻度を減らすことが望ましいとされています。未知語の登録作業の迅速化のためには未知語の自動的抽出手法が必要ですが、中国語は単語間にスペースをあけず、かつ漢字1字1字が意味を持つため、機械翻訳においては未知語の検出が難しくなります。ある単語が辞書に記載がなくても、1文字ずつに分解すれば辞書に記載があることが多く、そのように分解して翻訳した場合、訳は不自然なものとなることが多いものの、その不自然さの機械的な判定は困難なためです。そこで未知語の検出をどのように行うかを検討しました。

未知語は外来語に起因するものが多いことから外来語の中国語における表現方法を分析し、音訳と意訳の二つの手法にわけて検討しました。音訳語については一般に音訳に使用する漢字はある程度定まっているとされることから、音訳に使用されることが多い漢字をキーにして音訳語の未知語を抽出する自動抽出手法を検討しました。一方意訳語については一般的な自動抽出手法を確立することはできませんでした。また機械翻訳での検出が可能である未知語もあり、この場合の対応方法も検討しました。通常文献より未知語候補を抽出し、目視でチェックを行った後、誤抽出を除いた語に人手で対訳を付与する手法をとり、化学用語を多く検出することができました。

9. 文単位の特性の分析

機械翻訳は文単位で行うため、文の特定をいかに正確に行うかは機械翻訳結果を大きく左右します。そのため、中国公開特許公報の各部分の文の特性を分析しました。正しく文を特定することにより、機械翻訳結果の品質向上が期待できます。その結果、文末記号としては、「。」、「；」、「：」が候補文字として妥当であることが分かりました。さらに、特許文献には箇条書きが多く見られ、これを1単位の文として把握することも重要です。このような箇条書きを検出するキーとなる箇条書き項目の行頭記号（(1) や (a) など）と文末文字を決定することにより、ほとんどの箇条書き表現はカバーできることが分かりました。

10. 名詞句の分析

中国公開特許公報の大部分を構成する名詞句を、いかに正確に機械翻訳できるかが機械翻訳の精度向上に重要な意味をもつとの考えから、名詞句を分析し、文中にどれくらいの名詞句が含まれるか、かつそれがどのような合成規則による名詞句かを調査しました。調査の結果、中国公開特許公報の記載内容は、比較的単純な名詞句の構成が80～90%以上を占めることが分かりました。

さらに、IPC第8版A－Hセクションから、通常文献及びローカル文献を各々2件ずつランダムに選び、これらの機械翻訳結果と人間訳との比較を行い、そこに登場する名詞句に着目して機械翻訳結果の正誤を調査しました。その結果、5割が複合名詞、3割が専門用語の登録によって改善可能であるとの結論が得られました。

11. 分離前置詞の分析

中国語特有の表現に間に何かをはさむことができる離合詞があります。中でも、ひとつの前置詞として解釈できるものが前部分と後半部分の2つに分離して登場するものを、分離前置詞と呼びます。その間には名詞句などが挿入されるような使い方をします（例.「在～上」で「～の上に、～に関して」や「为～起见」で「～の点から見て」という意味になる前置詞）。中国語の構文解析に際し、この分離前置詞の検出は機械翻訳結果に大きな影響を与えます。そのため、本調査では、中国公開特許公報中にどのような分離前置詞がどのくらい使用されているか、またどのように使われているかを調査しました。

調査の結果、使用頻度の上位24種で、累積頻度の91%を占めることが判明し、中国公開特許公報において、分離前置詞の使われ方に一定の傾向を読み取ることができました。しかしながら、分離前置詞の使われ方は単純名詞以外の構造を持っているものが多く、さらに、分離前置詞の抽

出自体が困難な場合が多いため、機械翻訳にあたっては特別な処理が必要であると考えられます。

12. ローカル文献の特性

ローカル文献は、外国語訳が存在しない中国語特有の表現・単語が含まれている可能性があります。これを検証し、機械翻訳結果の品質向上に必要な対策を探りました。

未知語の分析で得た未知語の検出方法により、通常文献とローカル文献とを比較した結果、検出可能な範囲内での未知語の出現率がローカル文献では、通常文献の倍以上であることが判明しました。また、漢方薬など中国語固有の単語や、四文字熟語など漢字の組み合わせによる中国語特有の訳が付与しにくい表現も見られ、未知語対応はローカル文献のほうがより困難であることがわかりました。一方、他の分析結果においては通常文献とローカル文献との間には大きな差異はみられませんでした。

13. 中国特許文献の解析困難性

中国語は他の言語と比べても解析の困難度が高い言語と言われますが、特に中国特許文献としての解析困難性について、以下の課題別にご紹介します。

13-1. 中国語固有の文法的特性

(1) 形態素 (文を構成する最小の意味単位) の曖昧性

➢単語の切れ目が明確でない

問題点は以下のとおりです。

①文字種が数字・記号類の他はほとんど漢字であり、日本語のように漢字の他にひらがな、カタカナ、アルファベットといった異なる文字種がないため、明確な単語の切れ目となる情報がありません。

②かつ、中国語は表意文字としての性格から、1字1字がほとんど1単語としての意味を持っており、それが、後続する文字と組み合わせて1単語となるケースも非常に多いため、極論すれば形態素を決定するにも、文脈を理解しながらでなければ決定できません。

③さらに、外来語を取り込むのも漢字で行い、その1文字1文字がまた1単語の可能性を持つため、一層困難性を高めます。

➢多品詞語

中国語には多品詞語が多くあります。特に基本語彙の中に、動詞と前置詞 (例.「在」)、および動詞と名詞の多品詞を持つ単語が数多く存在します。

その結果、文の中にそれが1つでも存在すると、文全体の解析結果に多大の影響を及ぼして、翻訳結果が全く理解できないといった例も数多くあります。また、そのケースでは、訳順も大きく乱れて、原文と訳文の対応そのものが想定困難といった事態にもなります。特にその傾向は長文になればなるほど多いため、クレームやアブストラクトで顕著となります。

これについて、明確な対策は難しいですが、次のようなことが課題としてあげられます。

・動詞と名詞をもつ多品詞語に関し、特許文献ではほとんど使用されない品詞は削除する、あるいは、品詞として採用するときペナルティ指数を付与しておいて、滅多に使用されない品詞に関しては、ペナルティ指数を極端に高くしておいて、採用されにくいメカニズムを導入します。例として、「装置」という単語は名詞と動詞の品詞が一般的にはありますが、特許文献でこれが動詞として使用されることはまずないにもかかわらず、動詞として解釈されているケースも見受けられました。このようなものは、特許文献翻訳では削除しておいたほうがよいと思われます。

13-2. 外来語の曖昧性

外来語は未知語の大きな要因となっています。特に外来語は、音訳によって当て字を使用することが多く、辞書に登録されていない場合は、それが未知語として認識されればいいのですが、中国語では一字一字が意味を持つ場合が多いため、まったく別の単語として認識される結果、全く意味のとおらない誤訳の原因となります。また、外来語の取り込みは音訳もあり意訳もあり、いずれも一定のルールに基づいて行われるわけではないため、表記のゆれの原因にもなり、一層問題を拡大します。

一般に未登録の外来語が文中にあったとして、それを類推するのは困難です。「8. 未知語の分析」でご紹介したように、調査では音訳専用に用いられやすい文字をキーとして、その前後をさかのぼって外来語を類推する手法が提案されています。

外来語に対する対策としては、ともかく外来語を迅速に発見して迅速に登録していくというのが最善です。初期的な作業としては、日本語辞書中のカタカナ語、および英語辞書中の固有名詞にあたる単語をできる限り登録することが第一です。また、中国文献から前述の「音訳のみで構成される未知語」を探して、それに対応する外来語を登録するのも有効であると思われます。その際、対応する中国語あるいは日本語を探す際に、出現文に対する対訳メモリを持っておくことが重要です。

この日中対訳メモリの自動化については、完全に対応するものを自動抽出するということは困難であっても、高確率で対応文を抽出する手法はあるため、そういう手法をさらに発展させて、ある程度の精度でできるだけ多くの対訳

メモリ候補を自動抽出することが望ましいと考えられます。対応する訳語候補をその対訳メモリから最終的には人間がチェックして抽出するわけですので、対訳メモリ自体は100％の精度でなくても、ある程度の精度をもっていれば、十分役に立つと思われます。

13-3. 係り受けの曖昧

係り受けの主要な形式は、
・名詞句が名詞句に掛るケース
・前置詞句が動詞に係るケース
・節が名詞に係るケース
です。名詞句が単一構造で、動詞も一つしかないような単純な構造であれば、係り受けとして曖昧性もなく正しい解析も可能ですが、名詞句が構造をもち、さらにそれが並列になっていると、名詞句と名詞句の係り受けも可能な解が何通りも考えられ、曖昧になってきます。前置詞句が動詞に係るケースでも、動詞が複数個あれば、どちらの動詞に係るか曖昧となります。さらにこれらが組み合わさると、いよいよ曖昧性は増します。こういうケースは特許文献では頻繁に現れますが、さらに形態素解析で、本来名詞として解釈すべき単語が動詞として解釈されると、翻訳文としては、一体何をどう解析したか見当もつかなくなり、原文との対比もほとんど不能となってしまいます。

14. 翻訳精度向上のための定型表現・名詞句の辞書化

中国特許文献を翻訳することには、種々の問題点、困難性が伴います。特に請求項や要約に出てくる長文を全体として正しく翻訳することは、非常に難しいといえます。その中で確実に精度をあげることができる対策として、下記は方法論としても効果からも有力です。

14-1. 定型表現の登録

定型文、定型パターン、定型フレーズは、特許文献に頻出する特許固有の表現であり、かつこれらは、きまった形式で訳すことから、これらを抽出して翻訳に供することは、もっとも効果的な方法です。

14-2. 名詞句の登録

上記定型表現と並んで、中国特許文献の翻訳精度を確実にあげる方法として、できるだけたくさんの名詞句を登録していくことが考えられます。文全体の翻訳は難しくとも、フレーズ単位でみれば、翻訳のブレは少ないので、文の中で圧倒的に多くの部分を占める名詞句を確実に辞書化していくのは理にかなっています。その結果、文全体としての解析精度も上がっていくと考えられます。

特許文献には、PCT出願文献や中国と日本で優先権出願されて、中国と日本の両方に共通する特許文献も多いので、それを利用して、できるだけツールを使いながら、半自動的に高速に低コストでかつ高精度の辞書を作成していく手法の洗練が望まれます。代表的手法としては、以下の通りです。

自動アラインメント→高頻度語の抽出→対応する訳語候補の抽出→訳付のワークシート作成→人手による訳付

15. まとめ

以上、平成21年度に実施した「中国公開特許公報の機械翻訳による日本語での提供に関する調査」の結果を中心に、中国公開特許公報を日本語に機械翻訳するにあたっての課題や問題点等をご紹介しました。

中国語の特許情報の重要性が、今後ますます高まっていくであろうことを考えれば、中国語の機械翻訳精度について、更なる向上が期待されます。

また、本稿でご紹介した内容が、機械翻訳の観点だけでなく、中国語自体のご理解を深めていただく上でもお役に立てば幸いです。

profile

船守　茉美（ふなもり　まみ）

2006年10月　特許庁入庁。出願支援課に配属
2007年　4月　国際課
2010年　4月より現職