特許情報における機械翻訳の活用について

特許庁の取り組みを中心に

特許庁　総務部普及支援課特許情報企画室長　井上　博之

PROFILE
1991 年特許庁入庁。社会基盤分野、光学分野の審査官、審判官のほか、電子計算機業務課、経済産業省産業技術政策課、メリーランド大学客員研究員、情報システム室等を経て、2011 年 1 月より現職。

1 はじめに

現在、経済や企業活動のグローバル化に伴い、世界の特許出願件数は増加傾向にあり、特に中国における特許出願件数の伸び率は顕著である。このような状況の中で、世界の特許出願件数に占める中国・韓国における特許出願件数の割合は、1985 年の 4.4%から 2008 年の 24.1%に増加しており、一方で、世界の特許出願件数に占める日本・米国における特許出願件数の割合は、同時期で 59.6%から 44.4%に減少している[1]。

このような現象は特許情報の観点から見れば、これまで主に英語や日本語が中心であった特許情報において、中国語、韓国語といった日本語以外の非英語言語の割合が増えていくことを意味している。英語は、日本人にとって、他言語に比較して理解可能な言語であるが、中国語、韓国語を原語で理解できる人材は限られており、今後は中国語文献、韓国語文献の内容をどのように理解し、どのように検索するかということが重要となってくる。

中国語、韓国語を理解するにあたって、翻訳を介せば日本語で理解することが可能であるが、特に中国の特許・実用新案出願件数は、今後、大幅に増大することが見込まれており、そのような年々増加する膨大な数の文献に対して、人手翻訳のみで対応することはコスト面から見てもリソース面から見ても限界がある。したがって、このような非英語言語の特許情報を日本語で理解、検索するために、機械翻訳を利用することに期待が集まるのは当然であるといえる。

このような特許情報における機械翻訳の活用に注目が集まりつつある状況に鑑み、本稿では、特許庁（JPO）が既に取り組んできた特許情報に関する機械翻訳の利用について振り返ると共に、外国特許庁における機械翻訳の取組みについて概観したい。また、特許庁（JPO）において、これまで実施してきた中国語、韓国語から日本語への機械翻訳に関する調査結果を紹介する。

2 特許電子図書館（IPDL）、高度産業財産ネットワーク（AIPN[2]）における機械翻訳の利用

特許庁（JPO）においては、2000 年 5 月から特許電子図書館（IPDL）を通じて、2004 年 10 月から高度産業財産ネットワーク（AIPN）を通じて、機械翻訳を利用して日本語の特許情報を英語で発信している。具体的には、特許電子図書館（IPDL）においては、特許、実用新案、意匠公報のテキスト部分について、機械翻訳により英語で理解することが可能となっている。高度産業財産ネットワーク（AIPN）においては、外国特許庁のみの利用であるが、特許庁（JPO）の審査情報について、機械翻訳により英語で理解することが可能である。

日本語から英語への機械翻訳の精度は、他言語に比較して、既に高いものであるが、特許庁（JPO）では、精度向上に向けて、定常的に辞書の増強を行っている。特許電子図書館（IPDL）、高度産業財産ネットワーク（AIPN）において、翻訳不可能な単語（未知語）を収

1　WORLD INTELLECTUAL PROPERTY INDICATORS 2010 WIPO

2　Advanced Industrial Property Network の略

集しユーザー辞書に登録することや、高度産業財産ネットワーク（AIPN）における外国特許庁からの誤訳についてのフィードバックを分析・反映することにより辞書を増強している。また、2009年3月には機械翻訳エンジンのバージョンアップを行い、翻訳ロジックの改良や専門用語・知的財産権用語、翻訳用例（翻訳メモリ）が増強された。

特許庁（JPO）が提供している日本語から英語への機械翻訳精度について、Googleにおける機械翻訳と比較してNIST値、BLEU値、共に優れているとの評価もあり[3]、上記したような翻訳精度向上に向けた取り組みは一定の効果を奏しているということができる。

3 外国特許庁における機械翻訳の活用

前章で述べたように特許庁（JPO）は、機械翻訳を活用することにより、特許情報を英語で発信してきたが、外国特許庁においても、同様に機械翻訳を活用することにより、特許情報が英語で提供されている。これは世界の共通語ともいえる英語で、自国の特許情報を発信することが、結果として、自国民の知的財産権を守ることに繋がるなど、メリットが大きいと多くの国が考えているということである。

ヨーロッパ特許庁（EPO）においては、エスパスネット（Espacenet）を通じて、ドイツ語、フランス語、スペイン語、イタリア語の公報等の特許情報を機械翻訳により英語にして利用することが可能である。また、中国国家知識産権局（SIPO）、韓国特許庁（KIPO）においても、それぞれ同様に中国語、韓国語の特許情報を機械翻訳により英語で提供している。

ここまで自国語の特許情報を英語で発信することを中心に紹介してきたが、一方で、逆転の発想ともいえる他国の特に非英語の特許情報を自国語で自国の審査官や国民に提供するという考えもあるはずである。このような取り組みについては、ヨーロッパ特許庁（EPO）が早くから取り組みを開始しているようである。ヨーロッパ特許庁（EPO）は、非英語言語を自国語とする多くの国を抱えており、そもそも多言語間の翻訳についてのニーズが高かったものと思われる。現在、ヨーロッパ特許庁（EPO）は、28のヨーロッパ言語と日本語、中国語、韓国語、ロシア語の計32言語相互間で機械翻訳を用いることにより言語障壁を取り除くことを計画している。詳細な内容までは明らかにされていないが、ヨーロッパ特許庁（EPO）はGoogle社と連携し、Google社から機械翻訳のサービスを受けると同時に対訳コーパスをGoogle社に提供することにより機械翻訳精度を高めつつ、計画を進めていくようである[4]。

各国特許庁が機械翻訳を利用するにあたっては、どのように継続してそのサービスを維持していくか、また、どのようにして機械翻訳精度を向上していくかということは、コスト面など様々な問題が想定されるが、ヨーロッパ特許庁（EPO）とGoogle社の協力関係は、機械翻訳を活用する上での官民協力の一形態として、斬新なものであり、各国特許庁にとって大きな参考になるものである。

4 中国語、韓国語における機械翻訳の利用性

本章では、特許庁（JPO）が実施した中国語、韓国語から日本語への機械翻訳に関する調査結果を中心に紹介する。

4.1 中国語→日本語[5]

中国語から日本語への機械翻訳に関する調査では、実際の中国公開特許公報テキストデータを用いて、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、対応策の提案を行った。機械翻訳するにあたって想定される問題点として、いくつか指摘されているが、ここで

3 http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=142066

4 http://www.epo.org/news-issues/news/2011/20110324.html

5 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（平成22年2月）

は未知語となる可能性が高い異表記・表記のゆれと中国語の解析困難性について紹介する。

異表記･表記のゆれがある場合､それらのバリエーションが網羅されていないと未知語あるいは誤訳になる可能性が高くなる。異表記・表記のゆれは、外来語や組織名を中国語で表現する場合に起こりやすい。これは中国語における意訳[6]､音訳[7]の概念による部分が大きいようである。代表的な事例を以下に示す。

表１　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)。

表２　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)

表３　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

表４　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

6　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

7　漢字の音（読み）を使って原語の発音を表現する方法

また、中国語の解析困難性については、以前から中国語は他の言語と比較しても解析の困難度が高いと言われており、その原因は中国語固有の文法的特性によるところが大きい。具体的には、中国語は文字種が数字・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ないこと。かつ、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（文を構成する最小の意味単位）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ないなど、形態素が曖昧であり、単語の切れ目が明確でない。

さらに、中国語において外来語の取り込みは漢字で行っており、それらの外来語が辞書に登録されていない場合は、中国語では１字１字が意味を持つ場合が多いため、全く別の単語として認識される可能性があり、解析において一層困難性を高める。

また、中国語には多品詞語が多くあり、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する。その結果、文の中にそのような語が１つでも存在すると、文全体の解析結果に多大な影響を及ぼして、翻訳結果が全く理解できなくなる事例が数多くあり、そのようなケースでは、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やすく、特にその傾向は長文になればなるほど多いため、特許文献のクレームやアブストラクトで顕著となる旨報告されている。

4.2　韓国語→日本語[8]

韓国語から日本語への機械翻訳に関する調査では、中国語の場合と同様に、実際の韓国公開特許公報テキストデータを用いて、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、対応策の提案を行った。

韓国語の場合は、同音異義語が多いなどの問題が指摘

8　韓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（平成22年３月）

されているものの[9]、基本的には文法特性が日本語と非常に類似しており、ほぼ同じ語順で記述されるため、他の言語対で見られるような、原文と訳文の対応が取れなくなるということが避けられる。したがって、未知語や文の大半を占める名詞句などを登録していけば、機械翻訳のレベルとしては、かなりの精度が達成できるようである。

5 最後に

特許情報における機械翻訳の利用ニーズは高まっており、特に中国語から日本語への機械翻訳の利用については、喫緊の対応が必要である。

その一方で、中国語から日本語への機械翻訳の精度は未だ充分とはいえず、いかに向上していくかということが最大の課題といえる。課題解決の方法の一つとしては、未知語を登録するなど辞書の増強があげられ、今後、特許に特化した辞書の作成など検討する必要が出てくるものと思われる。

また、中国語の文法的特性に起因する解析困難性に対する対処は翻訳ロジックによる対応などが必要な場合もあり、より専門的な対応が必要で具体的にどのような対処が有効であるかは、今後、更なる議論が必要となるものと思われる。

9 例えば、「伝記」、「前期」、「前記」、「転機」、「電気」に相当する韓国語の読みは全て전기（チョンギ）となる。