内外国特許文献一括検索に向けて

Steps toward The Cross-Lingual Patent Retrieval System

特許庁　企画調査官　殿川　雅也

平成８年特許庁入庁。平成 27 年４月より現職、審査第一部調整課審査企画室に併任。

1 はじめに

平成 26 年に閣議決定された「『日本再興戦略』改訂 2014」及び知的財産戦略本部決定された「知的財産推進計画」において、「世界最速・最高品質の特許審査」の実現が目標として掲げられ、その実現のために内外国の先行技術文献調査を効率的に実施するための環境整備は重要である。

周知のとおり、内国特許文献と外国特許文献では、言語が異なることに加えて、内国特許文献に付与されている FI·F タームなどの分類·検索キー（以下「F タ－ム等」という）が外国特許文献の大部分には付与されておらず、国際特許分類である IPC や外国庁が付与した CPC 等の分類・検索キーを用いて検索する必要がある。したがって、内国特許文献及び外国特許文献を検索する際には、言語と分類・検索キーを切り換えて検索する必要があり、多くの場合、内国特許文献を調査した後に外国特許文献調査に移行する手順が採られている[1]。しかしながら、先行技術調査の対象となる外国特許文献は増大し続けており、先行技術調査の効率を高める施策はますます必要になっている。

当然のことながら、共通の言語と統一的な付与基準で付与された検索インデックスを用いて内外国特許文献を横断的に検索できれば効率がよい。それは、言語毎、発行国毎に検索手法を切り換えて実施することなく、引例が発見される蓋然性の高い文献集合から周辺領域の文献集合に順次調査範囲を拡大することができれば、検索手順の重複を省くことにより検索効率の向上を図ることが期待できるからである（図１）。これを早期に実現するために、機械翻訳技術及び分類付与技術といった基盤技術が重要となる。

これまで、言語横断検索、特許分類、特許機械翻訳について継続的に研究が行われ、知見の蓄積がなされてきた[2]。さらに近年は、電子計算機の能力が向上し、大量のデータを扱うことができるようになり、複雑なモデルを計算できるようになってきたことから、第３次 AI ブームとよばれる状況になっている。このような中で、上述した課題を解決する現実的なソリューションとして、これらの技術をどのようなかたちで適用できるのかを検討する事は重要である。

本稿では、特許庁内における検証[1]及び現在行っている調査事業を踏まえて、内外国特許文献一括検索実現に向けた最近の取組を紹介する。

1　ただし、現状においても、欧米和抄を利用することにより、検索範囲は限定的であるものの、和文による一括検索が可能である。また、機械翻訳全文や抄録翻訳文を対象に検索できる商用データベースを併用することも適宜行っている。

2　例えば、NTCIR (NII Testbeds and Community for Information access Research) プロジェクトでは、言語横断検索タスク、特許検索タスク（自動分類タスク）、特許機械翻訳タスク等のワークショップが開催されていた。特許機械翻訳の研究は、ワークショップ WAT (http://orchid.kuee.kyoto-u.ac.jp/WAT/) に承継されている。

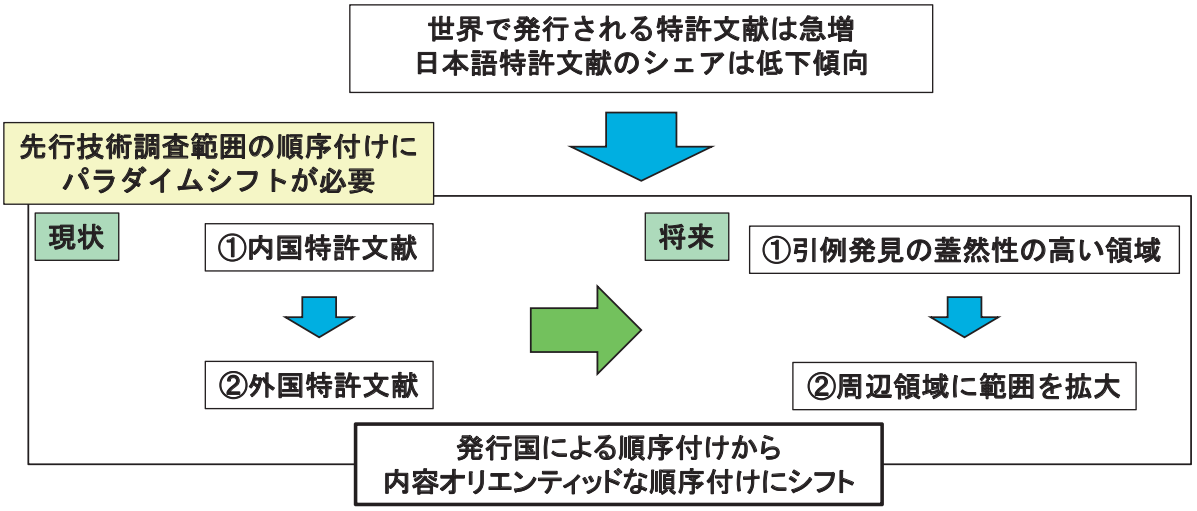

図１　発行国を意識しない先行技術調査へのシフト

2 特許機械翻訳

特許庁は、以前から多言語横断検索技術について検討を行ってきた[2] [3] [4]。その検討結果は、平成 26 年度にリリースされた中韓文献翻訳・検索システムに繋がっており、本システムは庁内外ユーザーの使用に供されているところ、機械翻訳文が外国特許文献の調査に有効であることが確かめられたといえる。

また、機械翻訳技術の活用を目指して、特許庁と国立研究開発法人情報通信研究機構（以下「NICT」）は、外国語特許文献の機械翻訳の必要性の高まりを受けて、平成 26 年に、中国語や ASEAN 言語等の機械翻訳の精度向上及び活用促進のための協力を行うことに合意し[3]、その後、平成 28 年に範囲を拡大した上で協力関係を継続している[4]。この協力を通じて、特許文献由来の高品質な対訳コーパスの普及とともに、特許庁での機械翻訳活用による特許審査の効率化、高品質化が期待されている。

そこで、この協力関係のひとつの成果として、大規模な対訳コーパスが提供されているので紹介する[5]。例えば、ここで提供されている英日対訳コーパスは約 3 億 5 千万文対を含むデータセットであり、特許についていえば、その規模は世界最大級のものである。現在、特許庁内でこれらの対訳コーパスに基づく翻訳エンジンを用いた機械翻訳ツールを検証中であるが[6]、検索に利用する機械翻訳文の品質として実用的な水準に達していると評価している。

さらに、最近では日独対訳コーパスの作成について協力を行ったところであり、庁内において検証を行っているところである。特許文献のパテントファミリーは、基本的に原文に忠実な直訳が記載されることから、文章の対応関係が比較的取りやすく、品質の高い対訳コーパスを作成するのに好適な素材として活用されている。ただし、発行国毎に明細書のフォーマットに差があるために、段落順が入れ替えられていたり、図の説明など異なる段落が差し込まれていることがあるため、対訳コーパス抽出の際に文章のアライメントをとるにあたり、発行国毎のフォーマットの違いを理解して作業しなければならない場合がある。そこで、日独対訳コーパスの抽出作

3　http://www.meti.go.jp/press/2014/07/20140728002/20140728002.html

4　https://www.nict.go.jp/info/topics/2016/04/160401-1.html

5　https://alaginrc.nict.go.jp/resources/jpo-info/jpo-list.html（本稿執筆時点では、研究目的に限定されている）

6　日英、日中、日韓、英日、中日、韓日に加え、後述する日独及び独日が対象である。

業では、文章のアライメントをとる際に行や段落のずれを吸収できるよう工夫し、得られた日独対訳コーパスを利用して、NICT の協力を得て翻訳エンジンを構築した。この翻訳エンジンを利用した日独及び独日機械翻訳についても庁内において検証中である。今後、対訳コーパスを作成する対象言語が拡大することが想定されることから、特許庁としては、発行国毎の特許文献のフォーマットの違いなど特許制度運用の知見やノウハウが必要な領域について協力できると考えている。

3 機械学習等を活用した分類付与

上述したとおり、外国特許文献の機械翻訳文を検索可能とすれば、異なる言語の特許文献を、言語を切り換えることなく横断的に検索できる。しかし、そのような状況であっても外国特許文献の大部分に F ターム等が付与されていない現状では、結局分類・検索キーを切り換えて検索しなければならない。これでは、内外国特許文献に対する一括的な検索の障害となるため、統一的な分類・検索キーの整備は急務である。

しかしながら、膨大な外国特許文献に対して内国特許文献と統一的な分類・検索キーを人手で付与することは現実的ではなく、分類付与技術の活用が必要となる。これまで、特許庁は、文献に分類・検索キーを機械的に付与する技術について検討した経緯があり [5]、その他の関連研究も継続的になされてきた（[6] の他多数）。

ここでの問題の解決策として、技術範囲が定義されたテーマのテーマコード [7] と、各々のテーマコードに対応する F ターム等を文献に機械付与することが考えられる。テーマコードは粗い検索キーと捉えることができ、FI・F タームはテーマコードより粒度が相当細かい検索キーと捉えることができる。したがって、まずテーマコードの機械的付与に取り組み、続いて、機械学習を FI・F ターム付与へ適用することの実現可能性について調査することとしている。

3.1 機械テーマコード付与

大部分の外国特許文献には F ターム等は付与されていないため、外国特許文献を検索する場合には、主に、IPC や CPC を利用している。

一方、内国特許文献には、FI を付与するとともに、上述したとおり FI が属するテーマのテーマコードが付与される。そして、全文検索を行う際には、テーマコードを用いて技術単位を指定し、その範囲内で全文検索することが通常行われている。

そこで、外国特許文献を対象に、テーマコードを機械付与することで、外国特許文献も内国特許文献と同様にテーマコードを指定したテキスト検索が可能となる。ここでは、おおまかにいって次の 2 つの手法を組み合わせた機械テーマコード推定方式を検討している。

（1）テーマに対応する CPC を予め定めておき、外国庁が外国特許文献に付与した CPC に基づき、文献に付与するテーマコードを推定

（2）日本語に機械翻訳された外国特許文献に対して、概念検索技術（特徴語の出現頻度等に基づく類似文献検索技術）を用いて全技術分野の内国特許文献から類似文献集合を抽出し、抽出された文献集合に付与されたテーマコードを統計的に解析して、外国特許文献に付与するテーマコードを推定

（1）の手法によればテーマと CPC の正相関が高い範囲では CPC に基づいてテーマコードを推定することができ、（2）の手法によれば、付与対象文献とテーマ内の文献との類似度からテーマコードを推定することができる。この2つの手法を組み合わせることで良好なテーマコード推定が期待できる。現在、この方式で外国特許文献にテーマコードを付与した場合の検索の有効性を庁内において検証しているところである。

3.2 外国特許文献への FI・F ターム付与の調査研究

前述のとおり、F ターム等は、テーマに複数含まれるものであり、テーマコードより一桁ないし二桁程度細かい検索キー（タグ）とみることができ、テーマコードより相当精緻に分類する必要がある。

内国特許文献には、F ターム等が人手で付与されてお

7 特許庁は、特許文献を技術範囲毎に区分して整備しており、各技術範囲を「テーマ」と呼んでいる。各テーマは、FI 範囲により技術範囲が定義され、5 桁のコードである「テーマコード」により特定される。現在約 2600 のテーマが存在し、そのうち約 1800 のテーマに属する特許文献に対して F タームが付与される。

り大量に存在しているため、機械学習を利用した分類付与手法と比較的親和性が高く、検索環境を効率的に整備することができるものと期待されるところ、現在、機械学習分野において技術進展が著しい状況にあるため、最新の機械学習手法を用いたＦターム等の分類付与技術の活用可能性を調査することは有意義である。したがって、平成28年度は、外国特許文献を対象にＦターム等の分類付与技術の活用可能性調査事業を実施しているところである。

本調査事業では、SVMをベースラインとして、深層学習（ディープラーニング）を含む複数のモデルを評価対象としており、最近の技術進展を反映した内容になる予定である。さらに、機械学習を活用した分類付与の実用化に向けて解決すべき課題を洗い出していく。

4 おわりに

本稿では、内外国特許文献一括検索に必要となる、機械翻訳技術及び分類技術について、特許庁における最近の状況を紹介した。これらの技術を有効的に組み合わせることができれば、内外国特許文献一括検索が実現できるだろう。

ただし、特許文献は、本質的に技術の進展とともに内容とその表現が変化していくものである。したがって、現在研究されている多くの機械翻訳技術ないし分類付与技術が表現に基づくものである以上、これらの技術を現場で運用する際には、技術の進展に伴う表現の変化に応じた学習を定常的に行えるようなフィードバックの仕組み[8]を運用に組み込む必要があると想定している。

したがって、これらの技術の導入にあたっては、品質維持向上のための学習プロセスを含む永続的に運用可能なサイクルを構築する必要がある。引き続き、検証を行うとともに現実的な導入工程を検討してまいりたい。

最後に、本稿に係る検討及び検証は、特許庁審査第一部調整課審査企画室次期検索システム検討WGメンバー諸氏によりなされたものであることを付記する。

8　現状では、新規の表現の裏にある概念を抽出することは人間が行う必要があり、人による何らかの品質チェック及び学習過程へのフィードバックの仕組みを組み込むことは必要だろう。

参考文献

[1] 殿川雅也、“これからの特実検索システムの探求”、特技懇誌、2016年1月29日、第280号、pp.31-38

[2] 特許庁、“「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”、2009年

[3] 特許庁、“「平成25年度 特許文献機械翻訳の品質評価手法に関する調査」調査報告書”、2014年

[4] 特許庁、“「平成27年度特許審査関連情報の日英機械翻訳文の品質評価に関する調査」報告書”、2016年

[5] 特許庁、“「審査関連情報を活用した次世代検索システム開発に向けた調査」調査報告書”、2009年

[6] 小林英司、“特許分類の自動推定に向けた取り組み　－機械学習による自動分類推定の課題と今後の展開－”、Japio YEAR BOOK 2015、2015年、pp.272-275