中韓文献の検索環境整備について

調整課審査システム企画班長　　後藤 昌夫

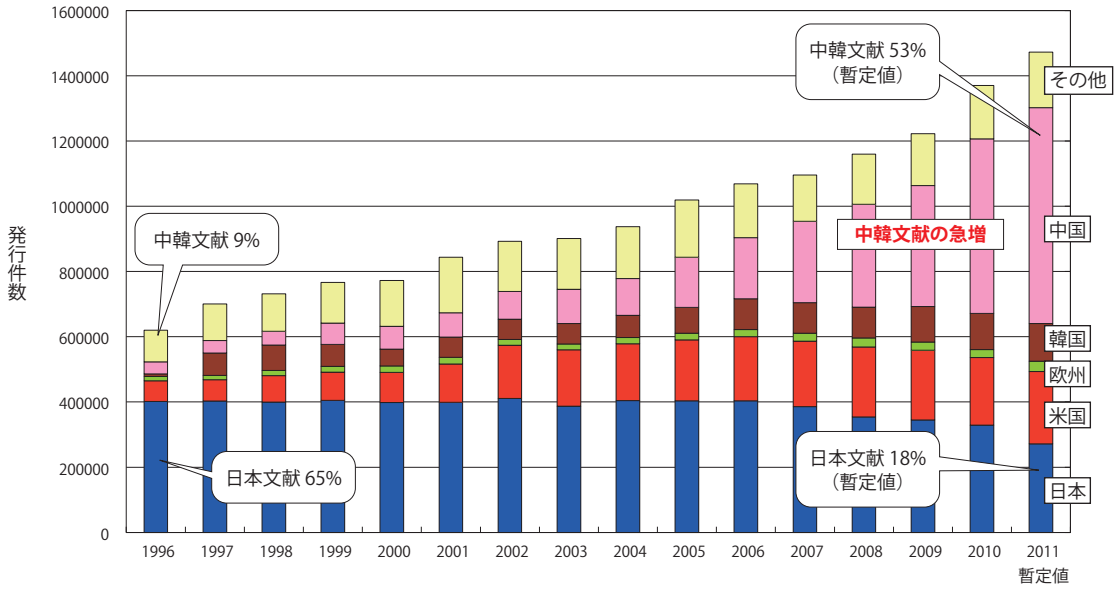
抄録

近年、世界の特許文献において、中韓文献をはじめとする外国特許文献の割合が急増しています。この状況に対応するため、中韓文献の検索環境整備を最優先で進めています。本稿では、審査のための検索環境整備の観点から、機械翻訳を利用した中韓文献の日本語全文検索システムを中心に、中韓文献の検索環境整備の５つの戦略について紹介します。

1. はじめに
～中韓文献をはじめとする外国特許文献の急増～

近年、世界の特許文献において、日本語以外の言語で記載された外国特許文献（実用新案を含む）の割合が急増しています。図１は、言語別に整理してパテントファミリーの重複を排除した、世界の特許文献の発行数のグラフです。例えば、日本文献と米国文献がパテントファミリーの関係にある場合、日本文献としてカウントしています。これによると、2001年から2011年までの10年間で、中国文献は約９倍に、韓国文献は約２倍に、欧米文献は約２倍に、それぞれ増加しています。

図２，３は、中国および韓国の特許と実用新案の出願件数の推移です。図１では横軸が公報の発行年であったのに対して、図２，３では横軸が出願年ですので、特許では約一年半、実用新案では半年～一年程度のずれがあります。中国においては、特許、実用新案ともに指数関数的な伸びを示しています。2011年の出願件数は、特許、実用新案ともに50万件を超えています。中国では、2015年に、特許の出願件数を75万件に、実用新案の出願件数を90万件にする計画がありますが、2011年までの出願件数の増加ペースを見ると、この計画を超える可能性もあると考え



（備考）世界で発行された特許文献（実用新案含む）を言語別に整理し、重複を排除したもの。複数の国に出願され、公開された同内容の特許文献について、日本語があるものは日本の特許としてカウント。日本語がない場合には、米国（英語）、欧州（英語、仏語、独語）、韓国（韓国 語）、中国（中国語）の順で該当する国・地域（言語）の特許文献としてカウント。2011年の発行件数は暫定値。
（資料）特許庁作成

図１　世界の特許文献数の推移

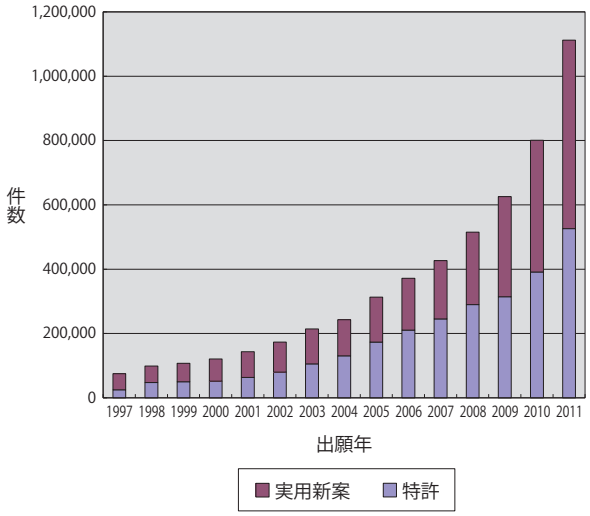


図2　中国（SIPO）における出願件数

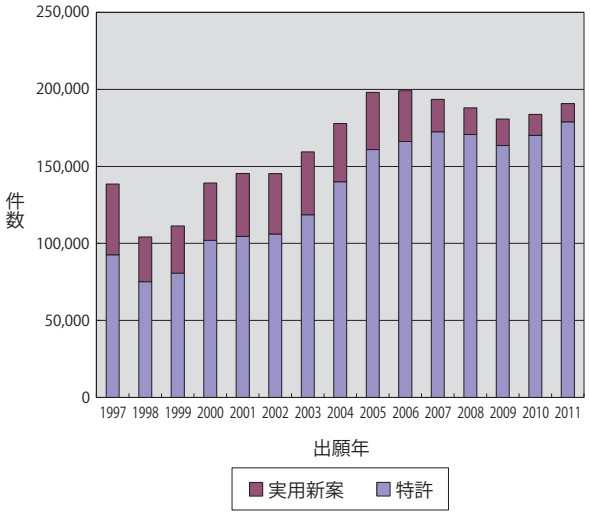


図3　韓国（KIPO）における出願件数

られます。また、韓国においては、特許について着実な出願件数の伸びが見られます。実用新案については減少傾向にあります。

このような状況において、世界で通用する安定した権利の設定を行うためには、外国特許文献についての的確かつ効率的な先行技術文献調査が不可欠であるため、これを可能とする検索環境の整備を進めています。中でも、中韓文献については、最優先で検索環境の整備を進めています。

本稿では、審査のための検索環境整備の観点から、機械翻訳を利用した中韓文献の日本語全文検索システムを中心に、中韓文献の検索環境整備の５つの戦略について紹介します。

2. 中韓文献の検索環境整備の５つの戦略

上述の通り増大する中韓文献を、容易に検索可能とするためには、翻訳と分類の２本柱について取組を行う必要があります。このため、図４のとおり、次の５つの戦略を推進しています。

・戦略１：中国実用和抄（機械翻訳）作成（2012年３月提供開始）
・戦略２：中国特許和抄（人手翻訳）作成（2013年３月提供開始）
・戦略３：中国特許FI/Fターム付与（2013年度提供開始予定）
・戦略４：中日辞書開発（2012年度～）
・戦略５：機械翻訳を利用した中韓文献の日本語全文検索システム開発（2015年１月リリース予定）

戦略１，２，４，５は翻訳に関する取組であり、戦略３は分類に関する取組です。これらを両輪として、中韓文献のアクセス容易化を推進します。以下において、各々の戦略について詳しく説明します。

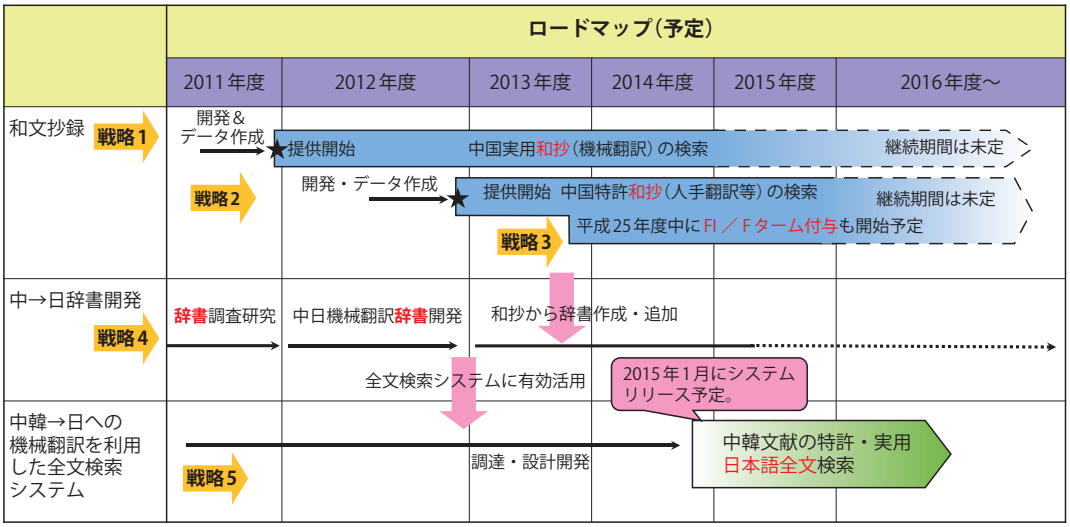


図4　中韓文献検索環境整備の５つの戦略

3. 戦略１：中国実用和抄（機械翻訳）作成

中国においては、無審査登録の実用新案権に基づき、訴訟が起こされるなど、リスクが高まっています。例えば、2009年4月には、実用新案権侵害訴訟に関し、フランス企業側が中国企業に1.5億元（約20億円）支払うことで和解しました。このような差し迫ったリスクへの対応として、中韓文献の中でも特にアクセスが困難であった中国実用新案について、日本語キーワードによる検索を可能とすべく、中国実用新案の和文抄録の作成が開始されました。この和文抄録は、2011年10月に行われた日中特許庁長官会合で合意が結ばれ、英文抄録データが利用可能となったことから、これを機械翻訳して作成したものです。

中国実用新案の和文抄録は、2012年3月にIPDLで検索サービスがリリースされました。2013年6月現在、170万件を超える和文抄録データを検索することが可能です。検索方法は、IPDLのトップページから「特許・実用新案検索」→「3.公報テキスト検索」とたどって、公報種別について「公開特許公報（公開、公表、再公表）」のチェックを外し、「中国実用新案機械翻訳和文抄録」にチェックを入れます。そして、検索キーワードの欄にキーワードを入力して検索ボタンをクリックすると検索できます。ヒットした中国実用新案の和文抄録は、PDF表示等が可能です。和文抄録を読んで、さらに明細書全文や図面を見たい場合には、Espacenetで見ることができます。Espacenetの「Smart search」欄に、公報番号が201435902以下であれば「CN（公報番号）Y」を、これより大きければ「CN（公報番号）U」を、入力してSearchボタンをクリックして下さい。

4. 戦略２：中国特許和抄（人手翻訳）作成

中国特許については、すでに、特実検索システムおよびIPDLにおいて、英語による要約（英文抄録）を提供しているところですが、2013年3月に日本語の要約（和文抄録）の提供を開始し、特実検索システムおよびIPDLにおいて検索可能になりました。この中国特許の和文抄録は、中国語の要約を人手により翻訳したものです。2014年3月までに、40万件を超える和文抄録データを作成する予定です。具体的な和文抄録の作成対象文献は、2010年および2011年に公開された中国特許出願公開公報であって、JP,US,EP,WOにパテントファミリーを有さない文献です。

IPDLにおいては、上記中国実用和抄と同様に中国特許和抄を検索することができます。「公報テキスト検索」画面において、「中国特許和文抄録」にチェックを入れることで、中国特許和抄の検索ができます。

特実検索システムにおいて中国特許和抄を検索するには、図6のとおり、まず、外国特許にチェックを入れ、発行国として「CN」を選びます（「全発行国」でも構いません）。

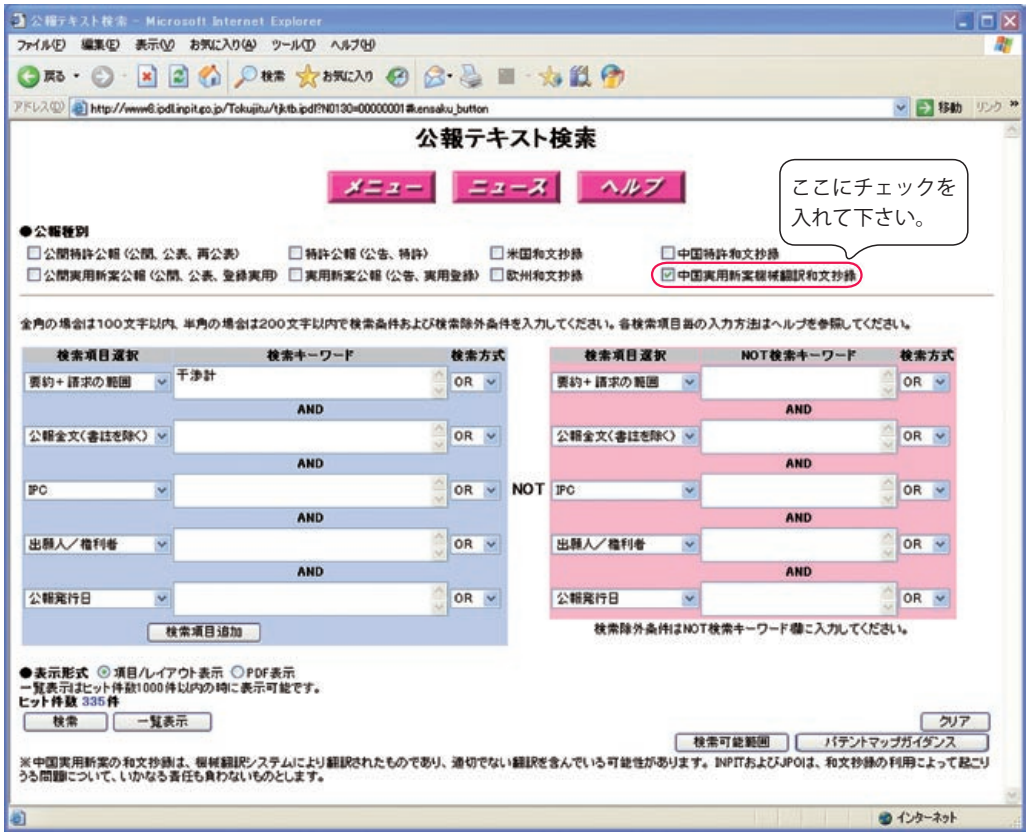


図5　IPDL公報テキスト検索画面

そして、検索指定が「和文」または「和文＋しおりメモ」になっていることを確認して、日本語でテキスト検索式（例：「光ファイバ/tx」）を入力して検索を行います。和文抄録の検索には、IPC検索（/ip）や後述のFI・Fターム検索を組み合わせることもできます。ヒットした中国特許文献をスクリーニングする際には、「表示」→「表示種別」→「二次文献」と設定すると、図7のとおり、和文抄録および英文抄録が表示されます。和文抄録と英文抄録の両方が存在する場合は、英文抄録が上に、和文抄録が下に表示されます。

中国特許和抄は、中国語の要約文から作成されるため、良質の対訳コーパス（中国語－日本語の文対）が得られることが期待されています。そこで、これを、後述の中日辞書作成に活用することを想定しています。

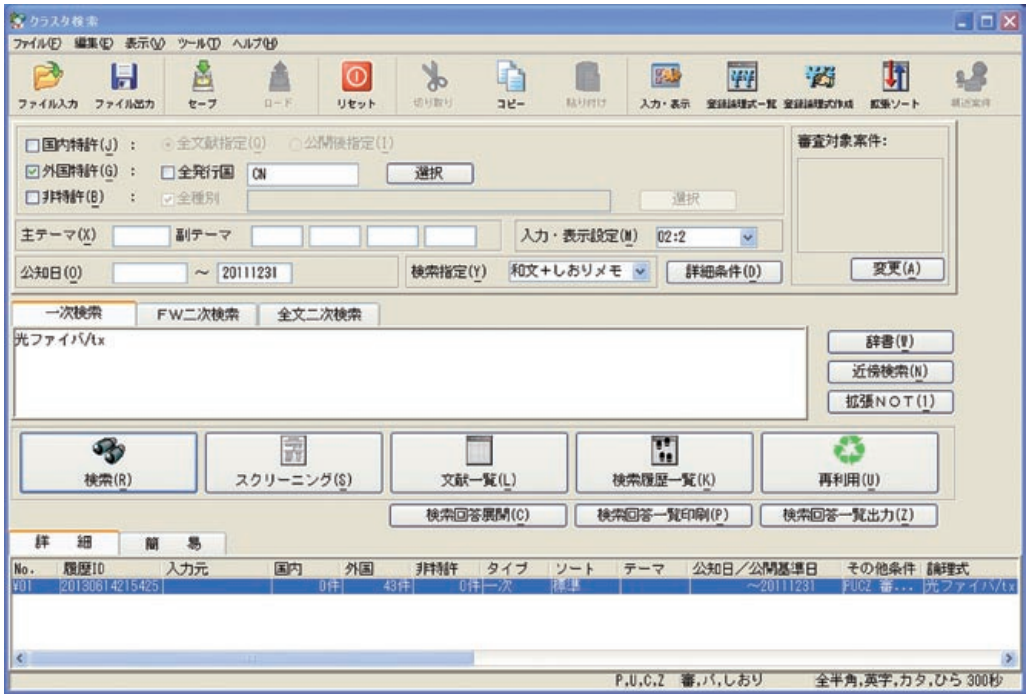


図6　特実検索システムにおける中国特許和抄の検索

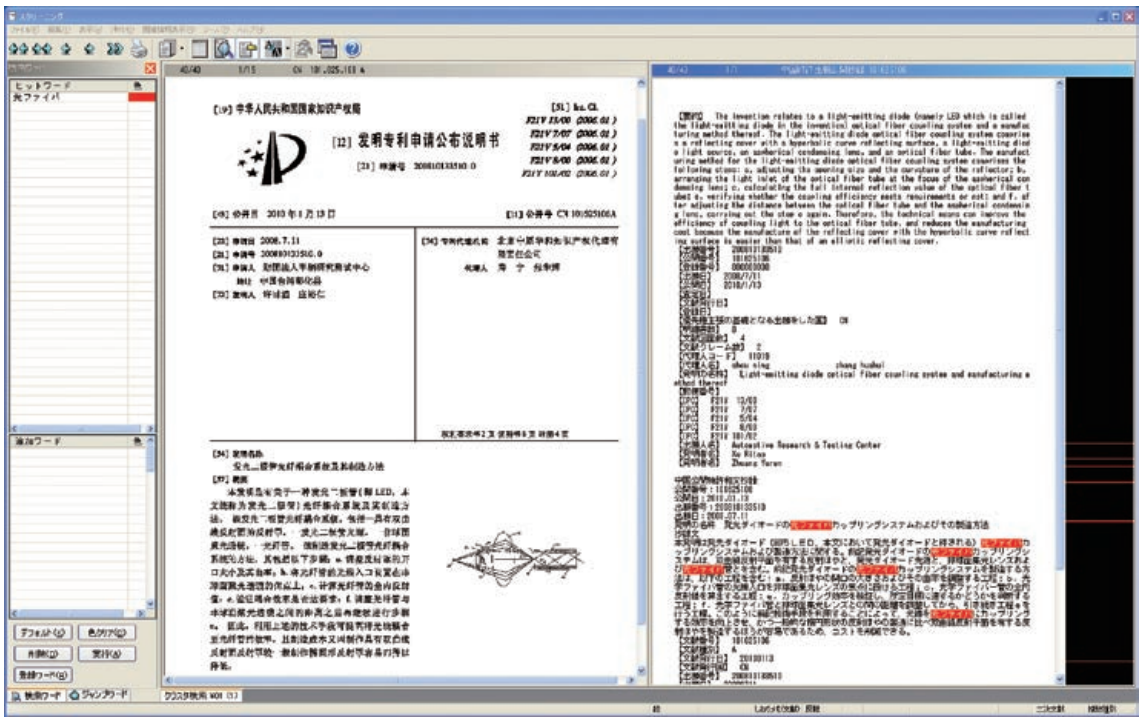


図7　中国特許和抄のスクリーニング画面

5. 戦略３：中国特許FI／Fターム付与

今年度から、中国特許文献へのアクセス環境の改善策として、中国特許文献にFI／Fタームが付与されます。現在、中国特許文献にはIPCが付与されていますが、分類項目が約7万項目と比較的少ないため、的確かつ効率的な検索には必ずしも十分であるといえません。これに対して、FIは約19万項目、Fタームは約38万項目あるため、これを中国特許文献に付与することで的確かつ効率的な検索に役立ちます。さらに、FI／Fタームの付与データは、中国国家知識産権局（SIPO）に対して提供を行うことで、知的財産分野における今後の中国とのより一層の国際協力と、世界に向けた日本分類のさらなる展開につながるものです。

今年度のFI／Fターム付与対象文献は、サーチニーズの高い技術分野について、2011年に公開された中国特許出願公開公報であって、JP,US,EP,WOにパテントファミリーを有さない文献であり、文献数は3～4万件です。

FI／Fタームが付与された中国特許文献は、特実検索システムおよびIPDLにおいて、FI／Fタームによる検索が可能となります。特実検索システムにおいては、「主テーマ」の欄にテーマコードを指定して検索を行うことができます。FI／Fタームが付与される文献は、上述の和文抄録も作成されますので、特実検索システムにおいては、和文抄録テキスト検索とFI／Fタームによる検索とを組み合わせることができます。

6. 戦略４：中日辞書開発

2011年度に、中国語から日本語への高品質な機械翻訳を実現するため、有効な翻訳辞書の開発方法を調査しました。その結果、調査の時点では中国語から日本語への機械翻訳の品質は、審査業務等で利用できるレベルに達していないことが明らかとなり、翻訳精度が不十分な原因として、中日辞書の規模が英日辞書の規模に比して小さい点が指摘されました。

そこで、中国特許文献の中日機械翻訳の精度向上に資することを目的として、2012年度に、実証調査の中で、特許文献で使用されている技術用語等について中日対訳辞書データの作成も行いました。当該辞書の語数に関しては、市販の中日機械翻訳ソフトウェアに搭載されている辞書の基本語が30万語程度であり、また、実績のある英日機械翻訳ソフトウェアにおいては通常100万語以上の辞書が備えられていることをふまえて、100万語となりました。

この中日辞書は、図8のとおり、次の4段階で作成されました。

(1) 2005年～2009年に公開された約100万件の中国特許文献を用いて、この中から、パテントファミリーの関係にある中国特許文献と日本特許文献との文献対（約26万件）を抽出します。
(2) 前記文献対から、自動文アライメントツールにより中日対訳コーパス（文対）を作成します。
(3) 中日対訳コーパスから辞書の見出し語、訳語の候補を機械的に抽出します。
(4) 各技術分野について専門知識を有する者が、機械的に抽出された訳語候補を確認して、中日辞書に登録します。

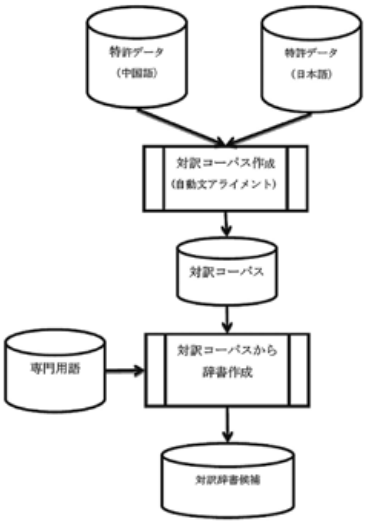


図8　中日辞書作成フロー

今年度以降も継続して中日対訳辞書データを作成することを予定しています。その際には、上述の中国特許文献の和文抄録データと原文の要約データとの文対を、辞書作成のために活用する予定です。これにより、最新の専門用語にも対応する見込みです。

7. 戦略５：機械翻訳を利用した中韓文献の日本語全文検索システム開発

(1) システムの概要と開発スケジュール

前述の翻訳辞書を有効に活用して、中韓文献の機械翻訳による日本語全文テキストを作成し、日本語による全文テキスト検索を可能にするシステムを2015年1月にリリースすることを目指しています。

本システムの概要は図9のとおりです。本システムにおいて翻訳および検索の対象となる文献は、中国特許出願公開公報、中国特許公報、中国実用新案公報、韓国特許出願公開公報、韓国特許公報、韓国実用新案公報であり、リリース時の文献数は約800万件（約10年分）を想定しています。また、リリース後は毎年約200万件の文献を追加していく想定です。中国と韓国の文献数の比率は4：1

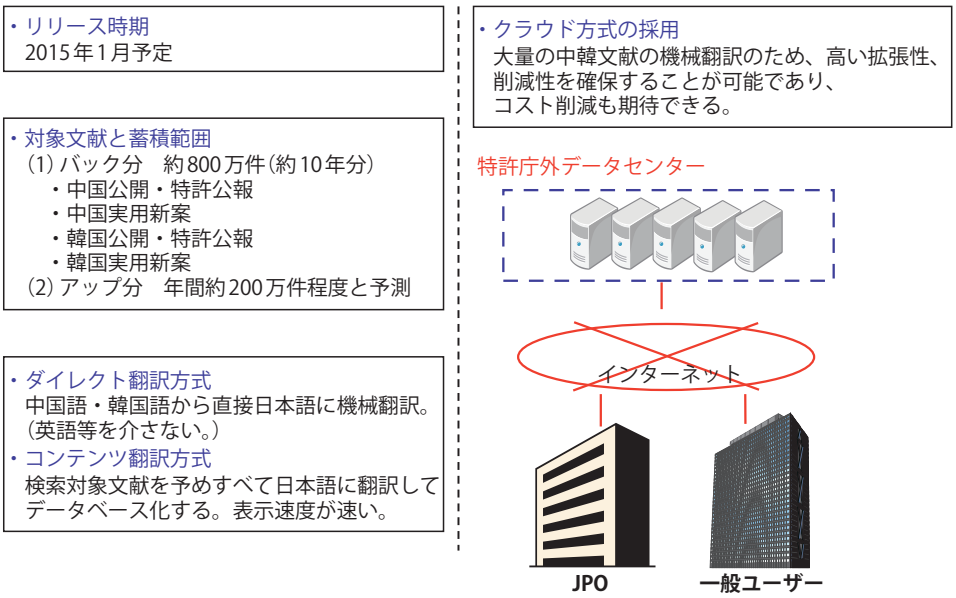


図9　システムの概要とリリース時期

程度となる見込みです。

翻訳方式としては、中国語・韓国語から直接日本語に機械翻訳する（英語等を介さない）ダイレクト翻訳方式を採用します。ドイツ語、フランス語等のように英語に近い欧州言語から日本語への機械翻訳においては、英語を介した方が訳質がよい場合もあります。これは、欧州言語と英語とは文法が似ていて機械翻訳しやすいこと、ならびに、欧州言語と英語との間、および、英語と日本語との間の機械翻訳技術がかなり成熟していて翻訳辞書も充実していることが原因であると考えられます。中国語・韓国語から日本語への翻訳においては、このようなメリットは少ないと考えられるため、ダイレクト翻訳方式となりました。特に、韓国語は日本語と文法が非常に近いため、ダイレクト翻訳方式による機械翻訳はかなり品質の高いものとなります。

また、翻訳検索方式としては、検索対象文献を予めすべて日本語に翻訳してデータベース化するコンテンツ翻訳方式と、日本語の検索キーワードを検索対象文書の言語（中国語、韓国語）に変換して原文のデータベースを検索するキーワード翻訳方式とがありますが、本システムではコンテンツ翻訳方式を採用します。コンテンツ翻訳方式の主なメリットとしてはスクリーニングにおいて翻訳文の表示が高速であることがあげられ、主なデメリットとしては検索精度が翻訳文を作成した時点での翻訳精度次第となることがあげられます。具体的には、翻訳文を作成した時点での未知語を含む文献は、その後に翻訳エンジンにその訳語が登録されたとしても、その訳語による検索でヒットさせることができません。一方、キーワード翻訳方式の主なメリットとしては訳語が増加するにつれて検索精度が向上することがあげられ、主なデメリットとしてはスクリーニングにおいて翻訳文をその場で生成（オンザフライ翻訳）するため翻訳文の表示が遅いことがあげられます。例えば、PCの翻訳ソフトでは一文献の翻訳に一分くらいかかることがあります。これらのメリット・デメリットを考慮した結果、コンテンツ翻訳方式となりました。コンテンツ翻訳方式のデメリットである検索精度の向上については、訳語の追加がある程度たまった段階で、すでに翻訳を行った文献について再翻訳を行うことを想定しています。

本システムは、クラウド上のサーバにおいて運用され、庁内および庁外からインターネットを介して利用する形態となる予定です。本システムでは、公開済みの文献に関する情報だけを扱うため機密性情報はないこと、また、大量の中韓文献の機械翻訳を行うには大量のコンピュータ資源が必要であることから、高い拡張性を確保することが可能であるとともにコスト削減も期待できるクラウド方式を採用することとなりました。また、本システムはインターネットを介して利用する形態ですので、各画面はInternet Explorer等のブラウザで表示されます。

想定される開発スケジュールとしては、今年度下半期にシステム設計を行うとともに、辞書・翻訳エンジンの整備を行います。そして、来年度には、システム開発およびテストを行うとともに、過去分の文献の機械翻訳を行い本検索システムに蓄積していきます。

(2) 現時点で想定する画面イメージ

現時点では次の図のような画面イメージを想定しています。ただし、今後の設計・開発で大幅な変更もあり得る点にご留意下さい。

図10は、検索画面の画面イメージです。IPDLと同じように、検索項目毎にキーワードを入力して検索を行う想定です。中国・韓国の特許・実用新案のうちどれを検索対象とするかを選択することができます（全てを選択することもできます。）。公知日や出願日の範囲を指定することもできます。また、近傍検索や拡張NOT検索も行えます。検索機能の詳細については次節で説明します。

図11は、文献照会画面の画面イメージです。（あくまで画面レイアウトのイメージです。翻訳文は架空のものであり、機械翻訳の品質を示すものではありません。）文献照会画面は左右2分割で表示され、機械翻訳文は画面左側に表示されます。文献送り等はショートカットキーで行うことができます。スクリーニング機能の詳細については次節で説明します。

図10　中韓文献翻訳・検索システムの検索画面イメージ

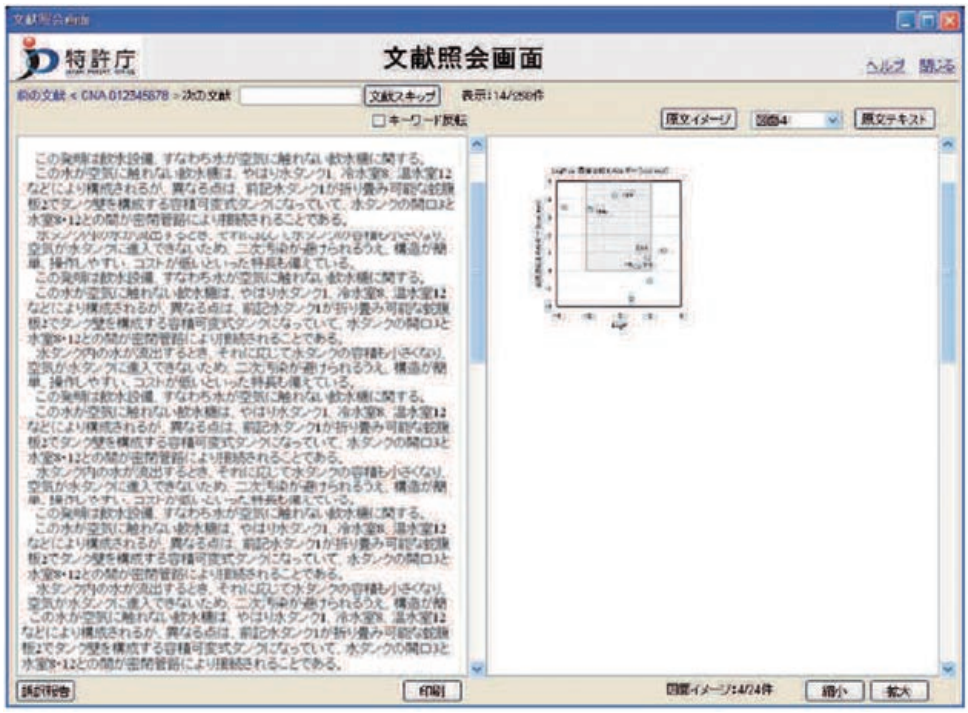


図11　中韓文献翻訳・検索システムの文献照会画面イメージ

(3) 本システムの機能要件および業務要件について

既に公表されている本システムの調達仕様書から、本システムの主な機能要件および業務要件をピックアップして説明します。この調達仕様書は、政府調達事例データベース（http://cyoutatujirei.e-gov.go.jp）において「中韓」のキーワードで検索するとダウンロードすることができます。

まず、文献検索に関する主な機能要件は下記の通りとなっています。

- 発行国、発行種別、文献種別で検索できること。
- 公知日（範囲指定含む）で検索できること。
- 出願日（範囲指定含む）で検索できること。
- 指定した文献の項目に応じて、中間一致検索ができること。
- AND、OR、NOTの検索ができること。
- 検索結果として、検索式に適合する発行国別の適合件数が取得できること。
- 検索式のサマリと検索結果（以下、総称して「検索式」という。）を一覧表示できること。また、その行に対して番号（以下、「検索番号」という。）を自動的に付与すること。
- 業務向け機能においては、文献集合リスト（CSV形式）のインポートができること。また、それら文献集合を検索式として保持できること。さらに、それら文献集合によるスクリーニングを（照会・表示）ができること。
- 業務向け機能においては、指定した検索式に適合する文献番号をエクスポートし、文献集合リストを生成できること。
- 入力されたカナ、英数字を全角、半角に異表記展開できること。
- 最大3語による近傍検索ができること。
- 拡張NOT検索ができること。

上述しましたとおり、発行国、発行種別、文献種別で検索できるとともに、日付範囲の指定ができます。また、AND、OR、NOTの検索ができます。審査官向けには、CSV形式の文献集合リストのインポート・エクスポート機能を使って特実検索システムとの間で連携することができます。（要件中の「業務向け機能」は、審査官向けの機能です。）例えば、特実検索システムで中国文献についてFI／Fタームによる検索を行ってヒットした文献集合をCSVファイルとしてエクスポートして、これを本システムにインポートして翻訳文の全文テキスト検索で絞り込むことができます。

また、3単語近傍検索や拡張NOT検索もできます。例えば、検索条件「ログ」、否定条件「プログラム」で拡張NOT検索する場合、文献中に「ログ」を含む単語が「プログラム」だけである場合はヒットせず、文献中に「ログ」を含む単語が「アナログ」と「プログラム」の2つがある場合はヒットします。

次に、文献スクリーニングに関する主な機能要件について紹介します。

- 照会画面は左右2分割で表示できること。
- 業務向け機能については、左側に翻訳文、右側に成型原文又は図面が表示できること。
- 翻訳文及び成型原文は、「【出願日】」や「【特許請求の範囲】」等、項目名が識別できる表示ができること。
- 翻訳文及び成型原文の文章中には、数式や表、化学式、塩基配列等の画像イメージが原文データで指定されていた位置と同じ位置に表示できること。
- 文献の適合件数を表示できること。また、表示中の文献が何件目にあたるかの整理番号を表示できること。
- 適合する文献集合内において、次文献又は前文献に移動できること。
- 照会中の図面を拡大・縮小表示できること。
- 誤訳報告画面を起動できること。
- ショートカットキーに割り当てられた動作が機能できること。（次文献、前文献等）
- 翻訳文において、検索式で指定した語句を検索項目毎に色を変えてハイライト表示できること。

照会画面は、特実検索システムのスクリーニング画面と同様に左右2分割で表示します。審査官向けには、翻訳文テキストと原文テキストとを対にして表示することができます。原文テキストをコピー＆ペーストすることで、インターネット上の他の機械翻訳サイトの翻訳結果と比較することもできます。次文献、前文献等の動作は、ショートカットキーで行うこともできます。そして、ヒットワードについては、複数色でハイライト表示することができます。また、スクリーニング中に誤訳を発見した際に、誤訳報告画面を立ち上げてこれを報告することができます。

次に、機械翻訳に関する主な機能要件について紹介します。

- 原文と訳文を対にした定型文や文型のパターン等をデータベース化して保持する翻訳メモリを利用できること。また、翻訳メモリを編集できること。
- 国際特許分類等を用いて技術分野を特定し、技術分野に応じた機械翻訳辞書を使用して、訳し分けができること。
- タイトル（発明の名称）、クレーム（特許請求の範囲）、発明の詳細な説明、それぞれを区別し、それぞれの区分に応じた適切な形で翻訳できること。

・数式・化学式・塩基配列が含まれる場合についても、文章の区切りを適切に判断し、翻訳ができること。
・有機化合物の特殊な表記に対応した翻訳ができること。
・機械翻訳辞書に登録されていない語句等を未知語として検出できること。

特許に特有の定型的な表現等について、原文と訳文の対を翻訳メモリとしてデータベース化して保持しておくことで、訳質を高めます。また、技術分野毎に機械翻訳辞書を使い分けることで、その技術分野にマッチした訳語を選択するようにします。また、機械翻訳の際に、辞書に登録されていない語句等を未知語として検出して、次で述べる訳質向上業務に役立てます。

最後に、訳質向上業務について紹介します。

受託者は、本調達で導入する翻訳エンジンにおいて、製品辞書が改版されたことにより、中韓文献サブシステムに適用することとなった場合、又は特許庁が毎年改版予定である中日機械翻訳辞書及び中日対訳コーパスが貸与された場合には、中韓文献翻訳サブシステムのユーザ辞書と重複する語句を調査するとともに、「2.5.2.3.(4) 集計業務」において集計された未知語や利用者及び誤訳パトロールにより報告された誤訳等の情報（以下、「未知語・誤訳情報等」という。）と併せて、ユーザ辞書や翻訳メモリ等への登録候補となる語句等を検討すること。

受託者は、機械翻訳で使用する製品辞書（基本辞書、専門用語辞書）及びユーザ辞書の改版、翻訳メモリ等への登録、前処理・後処理の改善、パラメータ調整等による翻訳エンジンの改善といった翻訳品質（以下、「訳質」という。）向上案等を含む訳質向上策を検討すること。

機械翻訳する際に発見された未知語や、利用者及び誤訳パトロールにより報告された誤訳等の情報に基づいて、辞書の改版や翻訳メモリ等への追加登録、翻訳エンジンのパラメータ調整等を行います。これにより翻訳エンジンが改善され、新しく出てきた専門用語等にも対応することができるようになります。

8. 最適化計画（改定版）における外国特許文献の検索環境整備について

今年3月に決定・公表された最適化計画（改定版）には外国特許文献の検索環境整備について次のように記載されています。

3.1 世界最高レベルの迅速かつ的確な権利の設定に不可欠なシステムの基盤整備

3.1.1 産業財産権制度を取りまく環境変化への対応

〈第Ⅰ期に実施する項目〉

（1）多言語翻訳機能を活用したグローバル化への対応

安定した権利の実現には、審査官が先行技術調査を行う際、世界の特許文献に占める割合が近年急速に増加している中国や韓国の特許文献を、網羅的効率的に調査することが必要である。しかし、それら文献の内容を原語のまま把握することは困難であることから、中国・韓国語特許文献の翻訳検索環境を整備する。さらに、外国語特許文献のうち英語特許文献については、先行技術文献としての利用価値が高いことから、翻訳検索環境を整備する。

具体的には、多言語文書検索（あるいは言語横断検索）と呼ばれる技術（日本語の検索ワードを検索対象文書の言語に変換するか、あるいは検索対象文書を予めすべて日本語に変換する等の処理を行い、検索ワードと検索対象文書の言語を統一した上で検索を行う方法）を導入するとともに、最新の国際特許分類等の最適な検索キーを用いて外国語特許文献と日本語特許文献の一括検索を可能にする。

ここで、「第Ⅰ期」は概ね5年以内を指します。第一段落には中国・韓国語特許文献の翻訳検索環境整備が謳われており、本稿で紹介した各戦略は最適化計画においても重要な位置づけとなっています。

そして、第二段落の後半に記載されるように、最終的には、審査のための検索システムにおいて外国語特許文献と日本語特許文献の一括検索を可能にすることを目指しています。「機械翻訳を利用した中韓文献の日本語全文検索システム」は特実検索システムと別システムとして2015年1月にリリースされますが、審査において前記一括検索を可能とするために、最終的には、日本語特許文献の全文テキストと外国語特許文献の機械翻訳全文テキストとを一つの検索システムに集約していくことを目指すべきと考えています。

profile

後藤 昌夫（ごとう まさお）

2001年4月　特許庁入庁（特許審査第一部光デバイス）
2005年4月　審査官昇任
2006年10月　調整課企画調査班調査係長
2007年10月　特許審査第一部計測
2010年7月　ボストン大学客員研究員
2012年1月より現職