中国文献へのアクセス向上施策
―中日機械翻訳システム開発環境の整備―

特許庁　総務部普及支援課特許情報企画室長　**森藤　淳志**

PROFILE　平成４年 特許庁入庁、機械工学分野の審査官、審判官等を経て、平成 24 年７月から現職。

特許庁　総務部普及支援課特許情報企画室調査第二係長　**稲葉　崇**

PROFILE　平成 18 年 特許庁入庁、通信分野の審査官を経て、平成 24 年１月から現職。

特許庁　総務部普及支援課特許情報企画室調査第二係　**船守　茉美**

PROFILE　平成 18 年 10 月特許庁入庁、出願支援課、国際課を経て、平成 22 年４月から現職。

1　はじめに

企業活動のグローバル化に伴い、全世界の特許出願は増加し、特許出願先の中心は従来の日米欧から日米欧中韓へとシフトしている[1]。特に、中国における特許の出願件数は急伸しており、2010 年に我が国を抜き、2011 年には米国を抜いて世界一位[2]となった。これにより、中国文献の十分な先行技術調査は世界で通用する安定した権利の確保、企業活動上のリスクの回避のために必須となっている。

英語文献については、従来から我が国審査官及び出願人が原語でその発明の内容を理解することが可能であると考えられていたが、急速に重要性を増してきた中国文献については原文のみから正確に内容を理解できる人材は限られているのが現状である。

このような情勢を受け、知的財産推進計画 2012 において、「中国語 … を始めとする外国語特許文献を日本語で検索可能な環境の整備を促進し、成果を出願人に提供する」ことが求められている[3]。

ここで、急増する中国文献の件数を考慮すると、全ての中国文献に対し、人手による翻訳を作成することは限られたリソース及び対応の迅速性の観点から困難であり、“機械翻訳を活用した”日本語による検索システムの開発が期待される。他方、機械翻訳のシステムの性能向上や機械翻訳の実施のためには、中国文献の機械可読なデータが必要であるが、中国文献のデータは種々の理由から市場に十分に流通していない状況にある。このことが、中国語の機械翻訳に関する研究開発の大きな課題となっている。

本稿では、上述の出願構造の変化や中国語の機械翻訳に係る課題を受け、米欧中韓特許庁及び世界知的所有権機関（WIPO）が実施する中国文献の検索・理解を目的とした取組の概況を紹介するとともに、日本国特許庁(JPO) による中日機械翻訳システム開発環境の整備等、中国文献へのアクセス性向上に資するための施策を紹介する。

1　国際知財戦略　～国際的な知的財産のインフラ整備に向けた具体的方策～　2011 年７月特許庁

2　特許行政年次報告書 2012 版「統計でみる知的財産動向」

3　知的財産推進計画 2012
http://www.kantei.go.jp/jp/singi/titeki2/kettei/chizaikeikaku2012.pdf

2 外国特許庁・機関における取組

2.1 各庁・機関における中国文献検索システムの現況

以下に、各庁・機関の機械翻訳を活用した中国文献検索システムの現況をまとめる。○は計画段階であること、◎は既に利用できること、－は計画の存在が確認されていないことを示す。

表1：各庁・機関の機械翻訳を活用した中国文献検索システムの現況

	内部利用	対外提供	対応言語	備考
JPO	○	○	内部利用·対外提供：中⇒日	
EPO	◎	○	内部利用：中⇒英 対外提供：中⇔EPOメンバー国の28言語、日、韓、露※1	対外提供：2014年末までに実現予定
USPTO	○	○	内部利用·対外提供：中⇒英	
KIPO	○	－	中⇒韓	
WIPO		◎	中⇒64言語（Google translate利用の場合）	PCT文献のみ

※1 EPO and Google remove language barriers from patent documentation http://www.epo.org/news-issues/news/2012/20120229.html

このように各庁・機関において中国特許文献への対応が行われており、その多くが、まだ計画段階である。このことが近年急速に件数を増してきた中国文献へのアクセス性確保が必須であるという統一認識と迅速な対応の必要性を示していると言える。以下に、各庁・機関の取組を具体的に紹介する。

2.2 欧州特許庁（EPO）

EPOは、2011年4月時点で、350万件の機械翻訳された中国文献の英語フルテキストを保有し、審査官が検索に利用している[4]。

さらに、EPOは、2010年11月にGoogleと特許文献の多言語の機械翻訳に関する協力について合意し、2011年3月に長期的な連携協定に合意した[5]。この多言語機械翻訳の対象言語には、中国語も含まれており、2012年2月にGoogleの翻訳技術を活用し一般に提供を開始した機械翻訳サービス「Patent Translate[6]」でも、2014年末までに中国語に対応したサービスの提供を予定している。

また、EPOは2010年9月[7]、2011年11月に中国国家知識産権局（SIPO）と機械翻訳に関する協力に合意しており[8]、協力関係を深化させていることが伺える。

2.3 米国特許商標庁（USPTO）

USPTOは、2012年6月に中国公開特許文献の機械翻訳サービスに関する情報提供要請[9]を行った[10]。中国特許文献の調査・検索のために、USPTOは中国特許文献原文テキストを提供し、事業者は職員には内部デー

4 http://documents.epo.org/projects/babylon/eponot.nsf/0/855ce511277876e1c125787700511a06/$FILE/schwander_documentation_matters_emw2011_en.pdf

5 EPO and Google break the language barrier for Europe's innovators http://www.epo.org/news-issues/news/2011/20110324.html

6 http://www.epo.org/searching/free/patent-translate.html

7 欧州特許庁、中国国家知識産権局と機械翻訳の協力に合意 http://www.jetro.go.jp/world/europe/ip/pdf/20100915.pdf

8 EPO and SIPO sign agreement on Chinese-English machine translation for patents http://www.epo.org/news-issues/news/2011/20111129.html

9 調達を行うための基礎資料として、外部業者に情報提供を求めること。

10 PROC1200208 Request for Information (RFI): Chinese Published Patent Document Machine Translation Services http://www.uspto.gov/about/vendor_info/current_acquisitions/PROC1200208_RFI_chinese_published_patent_doc_machine_translation_svcs.pdf

タベースを通じて、一般ユーザにはインターネットを通じて翻訳された中国特許文献を利用可能にする計画であることが上記要請から分かる。

2.4　韓国特許庁（KIPO）

KIPOは、2012年4月、審査官向けに中国特許文献の中韓翻訳サービスを新しく構築する計画を明らかにした。この計画は、2012年11月末に完了する予定である[11]。

2.5　世界知的所有権機関（WIPO）

WIPOの運営するPATENTSCOPEは9言語[12]で利用でき、検索結果にはGoogle translate, Microsoft translatorが利用できる。ただし、PATENTSCOPEで検索可能な中国文献はPCT出願に限られている。また、多言語検索（CLIR (Cross Language Information Retrieval)[13]機能により、検索用語を入力すると、検索用語の類義語とその外国語訳を検索用候補に挙げることで検索用語を拡張し、外国語で公開された文献を検索することができる。この機能において、中国語も検索用語拡張対象となっている。

加えて、WIPOは、発明の名称や要約のデータをもとに学習を行った統計的機械翻訳ツールであるTAPTA (Translation Assistant for Patent Titles and Abstracts)[14]の提供も開始している。英中間及び英仏間で双方向の翻訳が可能となっており[15]、先陣を切って多言語対応を進めている様子がうかがえる。

11 JETROソウル知的財産ニュース「機械翻訳サービス高度化事業」提案要請説明会
http://www.jetro-ipr.or.kr/

12 英語、ドイツ語、スペイン語、フランス語、日本語、韓国語、ポルトガル語、ロシア語、中国語

13 http://patentscope.wipo.int/search/clir/clir.jsp?interfaceLanguage=en

14 http://patentscope.wipo.int/translate/translate.jsf

15 COPPA, CLIR and TAPTA: three tools to assist in overcoming the Patent language barrier at WIPO
http://www.mt-archive.info/MTS-2011-Pouliquen.pdf

2.6　中国国家知識産権局（SIPO）

各庁から文献へのアクセスニーズの高いSIPO自身も、自国の特許情報の英語発信を行っている。SIPOの英語版ホームページでは、英語で検索できるPatent Searchを提供しており[16]、人手翻訳された要約のほか、全文の中英機械翻訳結果を得ることができる。そのほか、SIPOの関係機関である知識産権出版社（IPPH）が運営するCNIPR[17]はID, パスワード制による有料サービスではあるが、中国文献の英語によるテキスト検索が可能である。また、日本語による検索サービスについては、CNIPRのホームページ[18]に「日本語版CNIPR新発足」と記載されている。

3　JPOの取組

3.1　中国実用新案、中国特許の和文抄録作成

JPOにおける中国文献へのアクセス向上のロードマップを図1に示す[19]。

中国文献に対する差し迫ったリスクへの対応として、アクセス困難な中国実用新案について、英文抄録データから英日機械翻訳システムを用いて和文抄録データを作成し、2012年3月に特許電子図書館（IPDL）において、日本語キーワードによる検索サービスの提供を開始した。2012年度中には、過去10年分にあたる100万件超の和文抄録データが利用可能となり、また、新規発行分についても随時蓄積・提供する予定である。

中国実用新案に続いて、詳細な技術情報を把握する必要がある中国特許については、2012年度中に中国語の要約文の人手翻訳により、和文抄録データの作成を開始する。当該和文抄録データを用いて、随時IPDLにおいて検索サービスを提供開始する予定である。また、本

16 http://59.151.93.237/sipo_EN/search/tabSearch.do?method=init

17 http://english.cnipr.com/

18 http://japanese.cnipr.com/

19 特許行政年次報告書 2012年版
http://www.jpo.go.jp/cgi/link.cgi?url=/shiryou/toushin/nenji/nenpou2012_index.htm

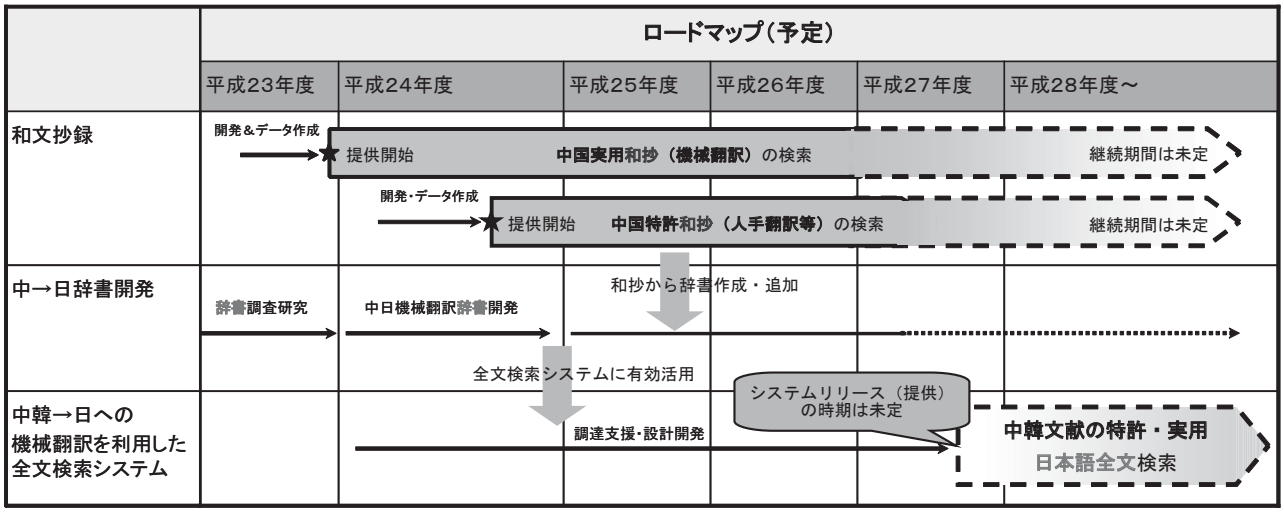

図１　中国文献へのアクセス向上のロードマップ

事業で作成される人手翻訳による和文抄録と、翻訳元の中国語の要約文は、良質の対訳コーパスとして、2013年度以降の辞書作成への活用等を検討している（後述3.3.1 も参照のこと）。

なお、IPDL に蓄積される和文抄録データについて、企業社内 DB や民間事業者 DB 等において活用できる形態での提供も実施する。

３．２ 外国文献検索システムの開発

和文抄録により差し迫ったリスクへ対応しつつ、知的財産推進計画2012における「中国語・・・を始めとする外国語特許文献を日本語で検索可能な環境の整備を促進」するとの指摘を踏まえ、中長期的には、中国文献全文の機械可読データを中国語から日本語へ機械翻訳する機能を備えた中国文献機械翻訳・検索システムの開発を視野に入れている。

また、同計画の「成果を出願人に提供する」との指摘も踏まえて、同システム開発の成果は、ユーザが活用できる形態での提供も検討してまいりたい。

３．３　中日機械翻訳システム開発環境の整備

３．３．１　中日対訳辞書、中日対訳コーパス作成

中国文献へのアクセス向上のための機械翻訳の活用に関して過去に JPO が実施してきた調査[20, 21, 22]により、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ている。それらの調査結果を踏まえて、2012 年度に中国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、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を実施する。

成果物たる中日対訳辞書データは、上記３．２で記載した外国文献検索システムにおける活用を想定している。

当該辞書整備事業について以下で紹介する。辞書整備事業の概要を図２に示す。辞書作成には、JPO が保有する、2005 年から 2009 年公開の約 100 万件の中国特許公報を用いる。この中から、ファミリー関係にある中国特許公報と日本特許公報との文献対（約26万件）を抽出し、自動文アライメントツールにより中日対

20 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（平成21年３月）

21 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）

22 特許文献の機械翻訳のための辞書整備に関する調査（平成24年２月）

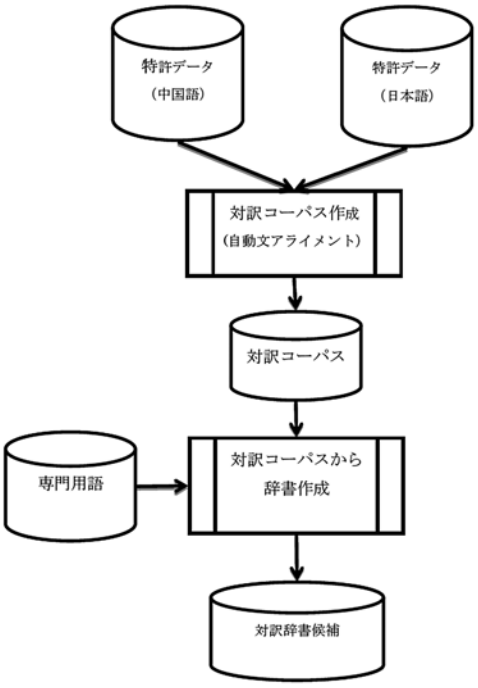

図２ パテントファミリーからの辞書作成

訳コーパスを作成する。作成する中日対訳コーパスのイメージを図３に示す。辞書作成は、当該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、その後、機械的に抽出された対訳辞書候補を、各技術分野に専門知識を有する確認要員により目視確認することで行う。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は 30 万語程度であり、又、実績のある英日機械翻訳ソフトウェアにおいては 100 万語以上の辞書が備えられているのが通常であることを踏まえ、本事業では、100 万語の中日対訳辞書データの作成を目指す。作成する中日対訳辞書のイメージを図４に示す。技術用語や専門用語の多くは名詞であると考えられるため、名詞を辞書作成対象とした。また、日本語でサ変動詞になるような用語についても、辞書作成対象に加える。

３.３.２　中日対訳辞書・コーパスの対外提供

機械翻訳システムは、高度な専門性、技術性を有するため、高精度の機械翻訳の実現にあたっては、機械翻訳メーカーや研究機関による研究開発成果を利用することが不可欠である。ここで、中日機械翻訳の研究開発には、

0.1000253624 ||| CNA1731511_JPA22006048058_des.txt ||| G10L15/06 ||| 本発明の構成要素、例えば、文字／音変換器１０５、公開語彙辞書１１０、混合言語HMMセット１１５、特徴抽出器１２５、ＡＳＲエンジン１３０、動的文法ネットワーク１３５などはすべて、コード読み取り専用メモリ（ＲＯＭ）６１２、文字読み取り専用メモリ（ＲＯＭ）６１４、ランダムアクセスメモリ（ＲＡＭ）６０４、スタティックメモリ６１６、およびＳＩＭカードの一つまたはそれ以上に、部分的または全体的に格納することができる。||| 本发明中诸如字模 - 发音转换器 105、开放词典 110、混合语言 HMM 集 115、特征提取器 125、ASR 引擎 130、和动态语法网络 135 这样的组件都可以部分或全部地存储在一个或多个代码只读存储器 (ROM)612、字符只读存储器 (ROM)614、随机存储器 (RAM)604、静态可编程存储器 616、和 SIM 卡中。

図３　中日対訳コーパスのイメージ※2, ※3

見出し語	訳語	品詞（見出し語）	品詞（訳語）
网络接口	ネットワークインターフェース	名詞	名詞
解扰	デスクランブルする	動詞	動詞

図４　中日対訳辞書のイメージ※4

※2 項目は左から、自動文アラインメント処理のスコア、特許文献の番号、特許分類、日本語文、中国語文となっている。
※3 本図はあくまでイメージであり、実際に作成される中日対訳コーパスの形式等と必ずしも同一ではない。
※4 本図はあくまでイメージであり、実際に作成される中日対訳辞書の形式等と必ずしも同一ではない。

専門用語や特異な表現等の分析のため、機械可読なデータの利用が不可欠である。例えば、対訳コーパスデータは、統計的機械翻訳における学習用のみならず、ルールベース機械翻訳等、他の機械翻訳方式でも翻訳精度向上のために有効である。

しかしながら、現状、中国文献データは十分に流通しておらず、中日機械翻訳辞書データや中日対訳コーパスデータの整備も不十分であり、中日翻訳研究開発の阻害となっている。

そのような状況に鑑みて、中日機械翻訳システム開発環境の整備の一環として、SIPO から JPO が入手した中国文献データを活用して作成した中日対訳辞書データや中日対訳コーパスデータを、研究機関等に提供することを検討してまいりたい。このデータ提供により、中日機械翻訳の研究開発が促進され、高精度の中日機械翻訳が早期に実現されることが期待される。

4 おわりに

以上、中国文献へのアクセス向上のための取組に関して、他庁等における現状、JPO における施策について見てきたが、各庁、機関において、翻訳言語の方向やその実施体制等は異なるものの、機械翻訳の活用という方向性は共通している。機械翻訳の活用に関する取組は、庁外の事業者との連携が不可欠であることから、引き続き、機械翻訳システム開発環境の整備に資する施策を進めてまいりたい。

また、本稿では中国文献への対応を中心に述べてきたが、外国語特許文献の割合が増加する中、他の言語の文献に対するアクセス向上のためのアプローチについても、今後検討していく必要がある。次なるターゲット言語はどれか、また、その言語の検索環境を整備するために官民で協力できることは何かについて、ユーザからのご意見も踏まえつつ企画立案し、適時に対応してまいりたい。