多言語横断検索技術について

株式会社東芝　研究開発センター　知識メディアラボラトリー　住田　一男
東芝ソリューション株式会社　特許ソリューション事業部　特許ソリューション第一部　樽井　伸司

1 はじめに

昨今の中国、韓国の出願件数は年々増加の傾向にある。中国、韓国の出願件数はすでに欧州での出願件数を上回っており（図1参照）、今後はさらに増加するものと予測される。

特に中国は2005年を境に韓国の出願件数を追い抜き、さらに急増している状況が窺える（図2参照）。また、出願件数のみならず、中国、韓国ともに技術進歩がめざましく、その特許文献の技術的水準も向上してきており、先行技術文献としても重要性は増す一方である。さらに、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たことも、中国、韓国の特許文献の先行技術調査の要請に拍車をかけている。

このような背景の下、日本国特許庁において、中国語、韓国語の検索環境の整備が緊急の課題として浮上してきた。一方中国語、韓国語を理解する審査官が英語に比べて極端に少なく、日本語で外国文献（中国文献、韓国文献）を検索する技術（多言語横断検索技術）に対する期待が高まっている。

本稿では、まず「多言語横断検索技術」の概要を説明した後、「審査利用へ向けた目指すべきゴール」を定義し、それを実現する上での「現状の課題と解決策（短期的視点）」を記載する。また、その解決策の有効性調査にも繋がる今年度実施中の「調査業務」及び、調査

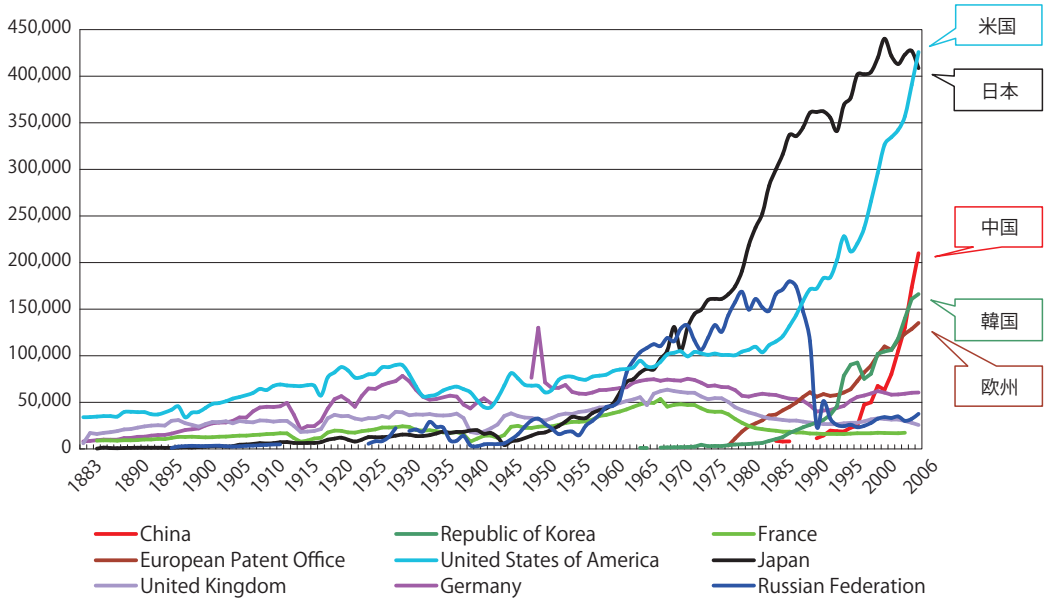

図1　世界の特許出願件数の推移（1883年～2006年）
出典：http://www.wipo.int/ipstats/en/statistics/patents/patent_report_2007.html

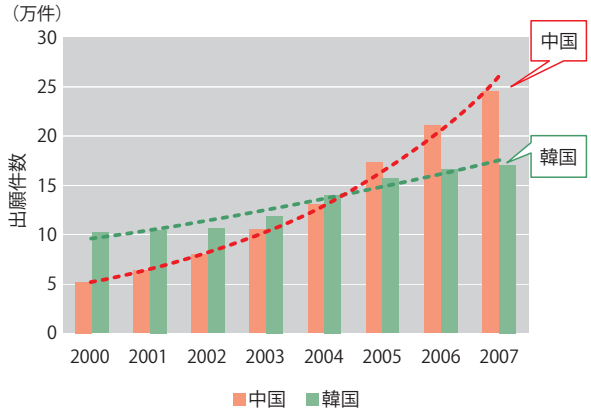

図2　中国、韓国への特許出願の推移
http://www.sipo.gov.cn/sipo_English/statistics/200706/t20070611_174616.htm
http://www.kipo.go.kr/kpo/route/FileDown.jsp?path=/upload/efile/&fn1=Applications.xls&fn2=Applications.xls

用に構築した「検証システム」を利用イメージも踏まえて紹介する。最後に長期的な視野から「今後の展望」について言及する。

2 多言語横断検索技術とは

多言語横断検索（multi-language information retrieval）技術は、母国語で検索条件を与えることによって、複数の言語で記載された文書を、言語を意識することなく検索可能にする技術である。

本技術は例えば、日本語のキーワード、文章などの検索条件（検索質問）を入力して、英語や中国語、韓国語などの外国語で記載されている特許文献を一括して検索する際に必要となる技術である。

検索対象の文書ごとに別システムで調査を行わなければならないとすると、利用者にとって負担が大きい。また、経済のグローバル化に伴い、特許情報などの科学技術文献がさまざまな国や言語で情報発信される傾向が近年強まりつつある。このため、複数の言語で記載された文書を、言語を意識することなく検索可能にすることが必要となる。

多言語横断検索技術は、情報検索技術と機械翻訳技術とを組み合わせ、融合させた検索技術であると言える。これまで様々なアプローチの情報検索技術や機械翻訳技術の適用が試みられてきた。

例えば、情報検索技術には、検索質問をキーワードで与えるキーワード検索や、文や文章で与える自然言語検索（概念検索とも呼ばれる）などが存在する。キーワード検索では、検索質問として入力したキーワードを含む文書はすべて検索され、検索もれは生じない。一方、自然言語検索では、検索質問を文章で表すことができ、適切なキーワードが思いつかない場合や、手元にある文書と類似する文書を検索する際に有効である。

また、翻訳技術の実現形態の観点で分類すると多言語横断検索は、検索質問を検索時に機械翻訳する「キーワード翻訳型」（逐次翻訳型）と、検索対象の全文書をあらかじめ機械翻訳しておきデータベースに登録しておく「コンテンツ翻訳型」（事前翻訳型）の2つの形態に大別できる。図3に「キーワード翻訳型」と「コンテンツ翻訳型」の概略構成を図示する。

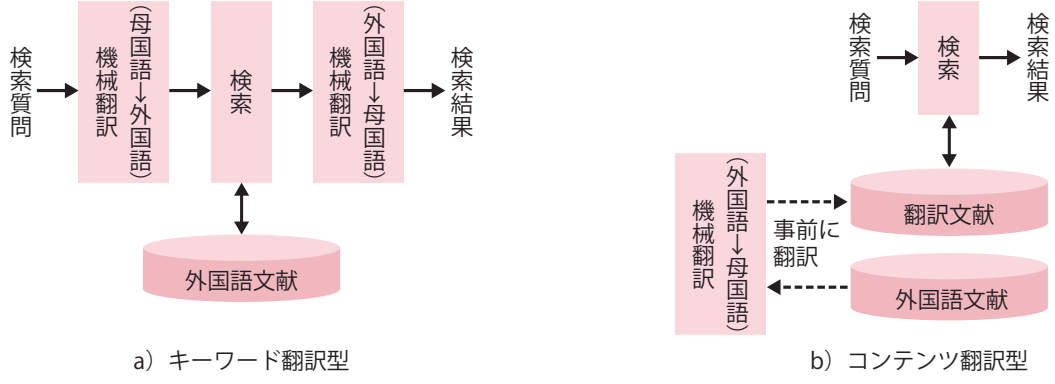

図3　多言語横断検索システムの概略構成

表1　キーワード翻訳型とコンテンツ翻訳型の比較

処理方式	観点	優位な点	不利な点
キーワード翻訳型システム	一般論	機械翻訳の訳語情報の更新に対して、システムの追従が容易（検索時に検索質問の翻訳を行うため）。	検索実行時に翻訳処理が必要なため、検索レスポンスの低下を招きかねない。文脈がないと正しく訳せないので、キーワード検索には向かない*1。
	審査業務	常に最新の翻訳辞書、翻訳システムの結果を利用できる。	スクリーニング時に、翻訳文の照会に時間がかかる。（翻訳処理が入るため）
コンテンツ翻訳型システム	一般論	検索エンジンは一つの言語に対応すればよく、検索エンジンの構成は単純になる。	データベースへの反映コストが大きいので、新語や訳語情報の更新に対して追従するのは困難。原文表示が必要な場合、翻訳結果とともに原文も蓄積しておく必要がある。
	審査業務	スクリーニング時に、JP文献と同等の性能で照会できる。（翻訳処理が入らないため）	翻訳辞書や翻訳システムのバージョンアップへの追随が難しく、訳質が悪いままで使わなければならない。（新規技術分野はその傾向が顕著）追随には、過去蓄積文献の再翻訳＋再蓄積が必要。

*1 … 例えば、“bank” という語は一語だけでは川の「土手」であるか「銀行」であるかが曖昧だが、“bank account” のように “account” とともに用いられたという文脈があれば、適切な訳「銀行」が得られる。中国語の事例では、同じ「トラック」であっても、自動車の「トラック」は「卡车」、競技場の「トラック」は「跑道」、磁気テープの「トラック」は「磁道」であるなどの訳語の曖昧性があり、文脈がなければ訳し分けすることができない。

これまでの研究事例からの知見を踏まえ、「キーワード翻訳型」と「コンテンツ翻訳型」のそれぞれの優位な点、不利な点を表1に示す。「キーワード翻訳型」「コンテンツ翻訳型」双方ともに、一長一短があり、利用形態を考慮して望ましい実現方法を選択すべきである。

3 審査利用へ向けた目指すべきゴール

「1 はじめに」で述べた通り、中国・韓国文献の重要度は高まっており、これらの特許文献を検索する必要性も増加している。但し、中国語・韓国語には言語の壁があり、そのまま明細書を理解する事は現実的に困難である。

従って、言語を意識せずに検索でき、日本語で「審査官が欲しい情報」を検索・照会できることが必要であることから、以下の目指すべきゴールが考えられる。

①日本語による外国文献検索

検索条件を日本語で入力して、引例候補となり得る外国文献を検索することを可能とする。

②シームレス検索の実現

検索対象文献の言語を意識することなく、複数言語の文献を横断的に検索することを可能とする。

③外国文献（中国語、韓国語）の日本語表示機能

中国語や韓国語等、理解できる審査官の少ない文献でも、明細書を含めた全体を日本語で確認できる。

4 現状の課題と解決策　～短期的な視点～

多言語横断検索は、翻訳技術と検索技術の両輪によって成り立っている。但し、現時点でも、それぞれ以下の様な課題があり、それぞれの解決策が考えられる。

なお、下記の解決策は何れも、検証システムに実装し、後述の調査業務（「5 調査業務の紹介」、「6 検証システムの紹介」）にて、その妥当性や効果の調査を進めているところである。

4.1 翻訳技術

4.1.1 自動辞書メンテナンス機能

(1) 課題 ～専門用語辞書のメンテナンスコスト増～

各単語の訳が如何に正しく得られているかが、機械翻訳ならびに検索精度に影響する。特許は分野毎の専門性が強く、一般用語辞書では不十分であるため、分野に応じた専門用語の登録は翻訳精度、検索精度を向上させるためには必須となる。但し、登録を全て人手で行うのはコスト増を招くことになる。

(2) 解決策

専門用語登録のために、ファミリー文献（日本から優先権主張によって出願された対応特許）を利用し、特許文献の自動文章解析を行い、用語と訳語のセットを自動抽出する。なお、自動抽出した用語と訳語に関して、人手によるチェックを経て、辞書登録を行う。

登録までを完全自動化すると、誤った訳を辞書登録してしまうため、最終的なチェックは人手で行う必要がある。

また、専門用語の種類や訳は、処理対象の文献の分野ごとに異なるため、このプロセスは分野ごとに行う。

例えば、IPC分類C07では、対応特許（申請号「CN02118417.8」の中国特許と出願番号「P2001-167873」の日本特許）における下記のそれぞれの文に関して、下線部が対応付けられることにより、「三甲胺」と「トリメチルアミン」が対訳として抽出されている。

中国特許（申请号：CN02118417.8）

用简便的蒸馏操作，在经济方面得到了降低能量消费的成本高纯度的三甲胺。

日本特許（出願番号：P2001-167873）

簡便な蒸留操作で、消費エネルギーコストが削減され経済的に高純度のトリメチルアミンが得られる。

この方法を用いることで、後述の「5 調査業務の紹介」で調査対象としているIPC分類のC07では、次のような専門用語と訳語の対が抽出することができた。

表2　C07での対訳抽出例

抽出された専門用語	訳語	従来の訳
三甲胺	トリメチルアミン	三の第一アミン
合钌	ルテニウム	を合わせる
高产率	高収率	率いると多収穫
最终结晶槽	最終晶析槽	最終の結晶溝
治疗用药	治療薬	服用を治療する

同様に、IPC分類H01Lでは、次のような専門用語と訳語の対を抽出することができた。

表3　H01Lでの対訳抽出例

抽出された専門用語	訳語	従来の訳
位线	ビット線	人の糸
引出端	リード	引き出して持つ
机架固定区域	フレーム固定領域	機体は区域を固定する
虚设单元	ダミーセル	単元を飾り物にする
小面	ファセット	小さな顔

このように、自動辞書メンテナンスを実現することで、専門用語辞書の効率的な拡充が可能になると考えらえる。

4.1.2 訳語候補展開機能

(1) 課題 ～訳語の曖昧性による検索漏れの発生～

キーワード翻訳型では、利用者が検索質問を入力し、それを中国語や韓国語に機械翻訳する。但し、表1の通り、各単語の訳語は一意には決まらず曖昧性が残る。このため、本来の訳語と異なる内容で検索してしまうことによる検索漏れが発生する可能性がある。

(2) 解決策

文献全体を日本語訳にするには、いずれかの訳語に決める必要があるが、多言語横断検索での利用を想定した場合、これらの訳語の候補を追加して検索することにより、検索時の漏れの低減が期待できる。これは、一つに決めた訳語が検索対象の文書に含まれる単語と一致しない場合でも、訳語候補の中には一致する単語が存在する可能性があるからである。キーワード翻訳型では、この訳語候補を検索語に追加して検索を行う機能、すなわち「訳語候補展開機能」が必要と考えらえられる。

例えば、日英翻訳では、「車載システム」は、in-vehicle system, on-board system, in-dash systemといった複数の訳語候補を機械翻訳の途中結果として保持しており、最終的にin-vehicle systemと訳出される。訳語候補を同義語と見なして検索語に追加して検索を行うことで、訳の食い違いによる検索もれが救われる可能性がある。

4.2 検索技術

4.2.1 自然言語検索機能

(1) 課題 ～検索ワード厳選の難しさ、検索結果の並び順の不備～

検索条件に利用する検索ワード厳選はノウハウが必要で初心者には難しい。特に、企業・大学等含めた庁外ユーザへの解放を視野に入れた場合、簡易な検索手法も必要になる。

また、検索結果の並び順を意識しない場合、例えば、関連ある文献が検索結果の末尾に出てくるなど、本来重要視すべき文献に辿りつくまでに時間が掛ってしまうという課題もある。

(2) 解決策

検索ワードではなく、文章（自然文）で検索条件を指定し、その内容に対する関連度（検索スコア）順に文書を順序付けて検索する「自然言語検索機能」によって解決する。自然言語検索は、文節単位に単語を切り出す形態素解析という技術を用いて、取り出された単語に基づく単語索引を利用して実現されている。自然言語検索の処理の流れは以下の通りである。

①検索質問（自然文）を形態素解析により単語に分割する。
②分割された単語の中から品詞により検索に利用する検索語を選択する。
③検索語と検索対象の文書を照合し、各文書の検索スコアを計算する。
④ 検索スコア順に文書を並べる。

検索スコアの計算は、検索質問から抽出された単語の文書中の出現頻度やその単語を含む文書数、文書の長さなどの統計情報に基づく方式で行っている。

このスコアは、大きく以下の3つのルールで決定している。

(1) 単語の出現文書数が少ないほど、大きくなる。

基本的には、検索質問中に含まれる単語と共通する単語を多く含む文書の検索スコアが大きくなり、その文書が上位にランキングされる。

しかし、文書をランキングする上で、すべての単語を等しく扱うわけではない。例えば、検索質問側のどの単語が重要であるかを判定するため、検索対象のデータベース中で少数の文書にしか含まれない単語を重要視する重み付けを行っている。これは、多くの文書に含まれる単語は、検索における弁別能力が低いという特性に基づいている。

(2) 単語の文書中の出現頻度が多いほど大きくなる。

文書側の単語については、その文書中の単語のうち、出現頻度が高い単語を重要視する重み付けを行っている。これは、ある文書で重要な単語はその文書で繰り返し用いられるという特性に基づいている。

(3) 文書長が短いほど大きくなる（例えば、同じ出現頻度なら短い文書が優先される）。

この単語の出現頻度は、文書が長くなると多くの単語の出現頻度が高くなる傾向がある。そこで、検索スコアの計算では、文書長の影響がでないように、文書長が短いほどスコアが高くなる様な補正を行っている。

但し、上記のスコア決定のルールでも問題がある。検索質問と文書とで共通している単語によって検索スコアを算出すると、検索質問の単語と同じ概念であるにも関わらず別の表記の単語が文書側で使われているような場合、その文書は検索結果の上位にはあがってこないという問題がある。このような言葉の揺らぎや同義語に対応する手法として、同義語辞書を整備するアプローチも存在するが、網羅的に同義語を整備することはコストや手間の問題がある。そこで、いわゆる連想検索と呼ばれる手法として、「擬似適合フィードバック（pseudo relevance feedback）」という方法がある。擬似適合フィードバックは、次のように処理を行う。

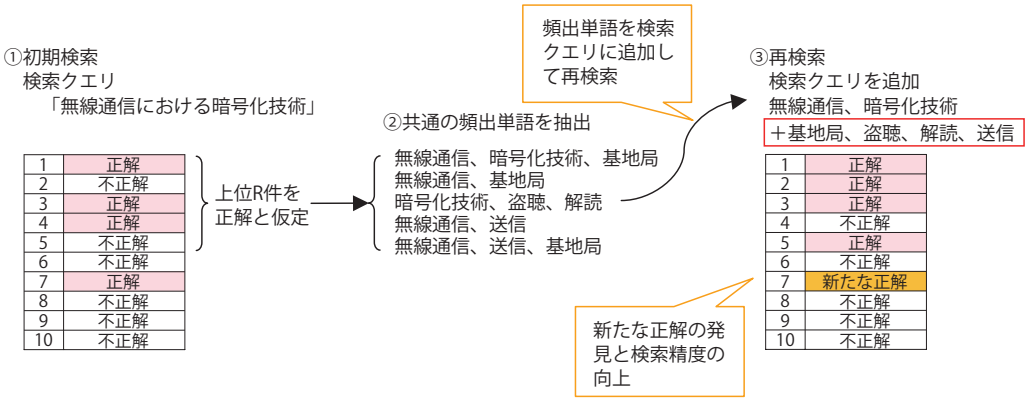

図4　疑似適合フィードバック

①入力された検索質問に基づき検索（初期検索）を行う。
②初期検索の結果の上位文書中の単語について検索語候補としての重みを計算し、その上位一定数の検索語候補を検索質問に追加する。
③②で拡張された検索質問を用いて再検索を行う。

擬似適合フィードバック機能によって、検索質問で明示されていない関連語が検索時に自動的に追加されるため、関連文書の検索に効果を発揮する。例えば、「無線通信における暗号化技術」という検索質問に対して、上記機能を適用すると初期検索で得られた文書中から「盗聴」、「送信」、「基地局」、「解読」といった関連語が抽出され、検索語として追加される。そして、これらの検索語を含めて再度検索を行うことにより、当初の検索質問に含まれていなかった「盗聴」や「送信」、「基地局」という関連語を含む文書も検索されることになる。

4.2.2 検索エンジンの各国語対応

(1) 課題 ～中韓対応の検索エンジンの用意～

キーワード翻訳型では、データベース内が中国語、韓国語等の原文で登録されている。このため、原文を検索するために、検索エンジンに中国語・韓国語の対応が必要となる。但し、中国語、韓国語に対応した検索エンジンは非常に少ないのが現状である。

(2) 解決策

今回の検証のために、日本語を対象とした自然言語検索機能をベースとして、試験的に中国語・韓国語にも拡張して対応した。中国語・韓国語への対応が必要な箇所は、基本的に単語索引を作成時に文書から単語を取り出す処理であり、その他の処理は共通である。

自然言語検索のため、中国語・韓国語のそれぞれで必要となる単語索引を作成するには、以下の2つの方法が考えられる。

・形態素解析：日本語と同様に、その言語に対応する形態素解析処理によって単語分割。
・Nグラム　：文字列をN文字単位（Nグラム）で切り出し、切り出した文字列を単語と見なして処理する（具体的には2文字単位（バイグラム））。

なお、形態素解析処理とは、辞書や文法を用いた処理によって、例えば、「東京都港区」という文字列に対して、「東京都」、「港区」という単語を取り出すことを言う。一方、バイグラムとは、「東京」、「京都」、「都港」、「港区」というように取り出した2文字ごとの系列のことである。バイグラムによる方法は、単語として本来存在しないものも部分文字列として出力するため、検索精度の低下の原因となる。加えて、形態素解析処理で得られる単語の種類は一定の数に収まるが、バイグラムは種類が多くなり単語索引の肥大化を招く危険性がある。実際のところ、今回、中国語解析について、バイグラムでの単語索引作成も試みたが、バイグラムの種類が増大し、単語索引の構築が困難であった。

一方、バイグラムを取り出す処理コストは、形態素解析のそれに比べて小さく、処理コストの観点からはバイグラム利用が望ましい。また、形態素解析処理に必要な辞書や文法が不要という特長もある。

「5 調査業務の紹介」で後述する今回の検証システムでは、中国語についてはバイグラムでの単語索引作成は困難であったため、形態素解析によって単語分割を行い、単語索引を作成している。一方、韓国語については、バイグラムで部分文字列を切り出し、その部分文字列を単語とみなして単語索引を作成している。

5 調査業務の紹介

執筆現時点（2008年11月末）、「多言語横断検索技術」を実際の審査で使えるか否かに関する調査を行っており、「4 現状の課題と解決策～短期的な視点～」で述べた解決策を実現した検証システムを構築して、その効果・妥当性確認、さらなる課題の抽出も行っている。

以下に現在実施中の調査業務について、その「目的と目標」「概要」「ユーザ検証の内容」の順に述べる。

5.1 調査業務の目的と目標

本調査業務は、平成26年1月稼動予定の「新検索システム」の要件定義を行うに当たって、新検索システム（特に、特許・実用新案検索システム）における、「多言語横断検索の①実現すべき機能の明確化、②技術的課題の抽出」を目的としている。

また、本調査の結果、新検索システムにおいて実現

表4　目標とする成果

項番	目標	内容
1	必要機能の洗い出し	多言語横断検索を審査業務に適用する場合の必要な機能を洗い出す。 例）自動辞書メンテナンス機能、訳語候補展開機能、シソーラス辞書による概念翻訳、検索結果の類似度順表示、等々。
2	基準値の提示	多言語横断検索の必要機能の基準値（参考基準値）を提示する。 例）多言語横断検索精度。
3	審査官の使用感	基準値等に表れない審査官の声をまとめ、必要機能に付加すべき事項を整理する。 例）アンケートによる使用感の声、意見の収集。
4	システムの実現性	①～③の結果、「新検索システム」の要件として、実装上の制約や現実的な内容かを評価。 例）　推奨アーキテクチャの提示、コスト面含めた実現性の評価。

すべき機能を明確にするために、平成21年3月末日までに、表4に記載の4点を、本調査業務のゴール（目標とする成果）としている。

5.2 調査概要

本調査業務では、「審査で使える多言語横断検索」即ち、「日本語で検索条件を入力することにより、中国文献、韓国文献の中から、引例候補を検索できること」を調査することを主眼としている。

従って、翻訳精度単体での調査を狙うのではなく、多言語横断検索の検索精度を高めることに調査のポイントを置いている。

また、調査は図5に記載した通り、以下の手順で進め、最終的に「審査で使える多言語横断検索」を調査する。

①中韓の特性分析（特許文献特性、言語特性）を実施し、その結果を多言語横断検索へ適用する。

②多言語横断検索技術そのものに加え、両輪である「翻訳技術」、「検索技術」も調査する。

③審査官（ユーザ）に用意したシステムを利用していただき、使用感を確認する。

5.3 ユーザ検証の内容

2008年11月上旬から2009年1月上旬まで、特許庁審査官に、「6 検証システム紹介」のシステムを使用していただき、審査官の目線から、多言語横断検索の効果・妥当性確認、課題の抽出を行っている。

特に、10名の審査官においては、擬似的な実体審査を実施していただき、類似文献を発見できるか否かを

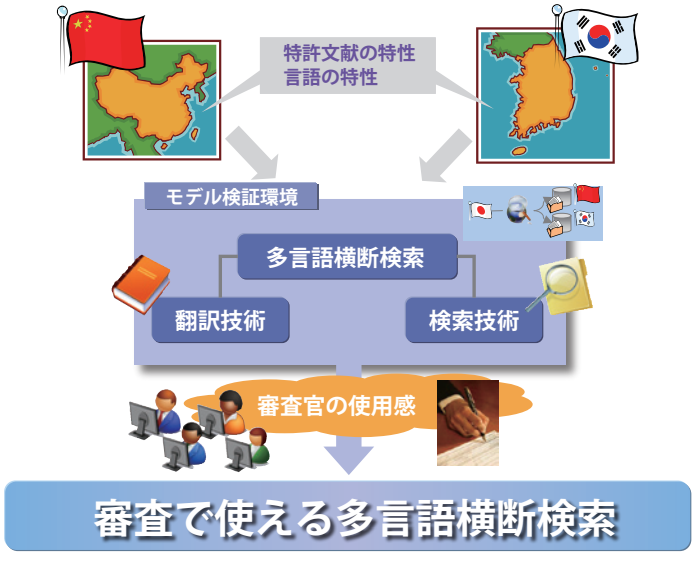

図5　調査概要

調査していただいている。

5.3.1 調査内容

ユーザ検証は、全4回に分けて実施している。以下の調査内容を、各検証フェーズにて調査し、効果や課題を確認する。

なお、検索精度の調査は、弊社においても、予め用意した正解がどの程度検索されるかという観点で実施している。検索条件には日本の公開公報を、正解には検索条件文献のファミリーである中国、韓国の公開公報を使用し、定量的に評価している。

(1) 処理方式の比較

1回目の検証と4回目の検証にて調査する。

表1の通り、「キーワード翻訳型」と「コンテンツ翻訳型」は、一長一短である。このため、「キーワード翻訳型」と「コンテンツ翻訳型」のそれぞれにおいて、同一の検索条件で検索し、検索精度の観点から比較を行い、多言語横断検索において効果のある方式を検証する。

(2) 自動辞書メンテナンス機能の有効性

2回目から4回目の検証を通して調査する。

検証2から4にかけて、「4 現状の課題と解決策　～短期的な視点～」にて述べた、自動メンテナンス方式にて段階的に分野別辞書の登録語数を増やす。同一の検索条件で検索し、検索精度の観点から比較を行い、自動辞書メンテナンス機能の有効性を検証する。

(3) 訳語候補展開機能の有効性

4回目の検証にて調査する。

キーワード翻訳型で検索する場合、検索条件を正確に翻訳することが重要である。このため、「4 現状の課題と解決策　～短期的な視点～」にて述べた、訳語候補展開機能と、文脈を意識して翻訳するため、正確に翻訳することができる自然言語検索にて検索し、検索精度の観点から比較を行い、訳語候補展開機能の有効性を比較する。

(4) 翻訳精度の確認

4回目の検証にて調査する。

同一文献において、①分野別専門辞書を使用せずに翻訳、②製品版の分野別専門辞書を使用して翻訳、③自動辞書メンテナンスにて作成した辞書を使用して翻訳、④シソーラス辞書を使用して翻訳した文献を審査官に確認していただき、翻訳精度の観点から比較を行う。

(5) IPC指定の有効性

2回目の検証にて調査する。

外国文献の検索において、IPC指定は必須と考えられる。但し、日本特許庁と諸外国の特許庁では、IPC付与の基準が異なるものも存在する。このため、IPC指定の有無のみが異なる、同一の検索条件で検索し、検索精度の観点から比較を行い、IPC指定の有効性を検証する。

5.3.2 調査対象分野

調査対象の分野は、表5の通りである。分野は、①中国、韓国国内での出願が多い分野、②日本から中国、韓国への出願件数が多い分野を優先的に選択した。

①の「中国、韓国国内で出願件数が多い分野」は、それぞれの国の主要な技術分野であると考えられる。また、②の「日本から中国、韓国への出願件数が多い分野」は、日本企業が中国、韓国を意識する分野であると考えられる。このため、以上の2点を兼ねる分野が、中国文献、韓国文献に対する先行技術調査のニーズが高い分野であると判断した。

表5　調査対象分野

言語	分野	(A)	(B)	タイトル
中国	A61K	1	586	医薬用、歯科用又は化粧用製剤
	C07C	—	417	非環式化合物または炭素環式化合物
	C07D	10	443	複素環式化合物
	H04L	2	1,110	デジタル情報の伝送
	H01L	4	3,775	半導体装置,他に属さない電気的固体装置
韓国	A61K	5	492	医薬用、歯科用又は化粧用製剤
	H01L	4	1,365	半導体装置,他に属さない電気的固体装置
	H04N	1	446	画像通信
	H04B	1	329	伝送
	G06F	2	405	電気的デジタルデータ処理

(A) 中韓国内での出願件数順位（2007年）
(B) 中韓国における、出願人が日本国籍の公開文献数

5.3.3 自動辞書メンテナンス対象分野

自動辞書メンテナンスは、中国文献のH01LとC07（C07C＋C07D）を対象として実施する。

H01Lは、対応特許が3500件と多く、且つ辞書としての分野の幅が狭い分野である。このため、用語と訳語のペアが発見し易く、一度に追加する用語数が多くなり、有効性の判断がつきやすいと考えられる。

また、C07は、対応特許が800件程度であるが、比較的分野の幅が狭い分野である。このため、H01Lほどではないが、比較的、追加する用語数が多くなり、有効性の判断がつきやすいと考えられる。

5.3.4 今後の予定

ユーザ検証終了後、2009年1月初旬から2月末にかけて、ユーザ検証の結果を評価・分析する。3月には、評価・分析結果を基に成果報告書を作成し、3月末の成果報告会にて、本調査業務の成果報告を行う。

6 検証システムの紹介

調査業務を遂行するに当たり、構築した検証システムについて「システム概念図」及び「利用イメージ」を説明する。

6.1 システム構成図

今回の調査を行うに当たり、用意したシステムのシステム概念図を図6に示す。

本システムの主要構成要素である「翻訳エンジン」、「データベース」、「ユーザインタフェース」の特徴は以下の通りである（図6の中で○数字にて記載）。

①自動翻訳システムには中韓それぞれ専用の翻訳エンジンを利用した。
　中日／日中翻訳エンジンは高度な検証を行うために製品「The 翻訳™サーバ」(注1) からコアロジック部を抜き出したものを利用した。韓日／日韓翻訳エンジンは「J−Server™」(注2) を利用した。

②検索処理にはXMLデータベースを使用し、原文のフォーマット変換なしにデータベースへの蓄積を行った。市販のXMLデータベース「TX1™」(注3) に試験的に中国語／韓国語対応を加えた。

③シームレス検索を可能とするユーザインタフェースを提供した。

（注1）（注3）：東芝ソリューション株式会社製
（注2）：株式会社高電社製

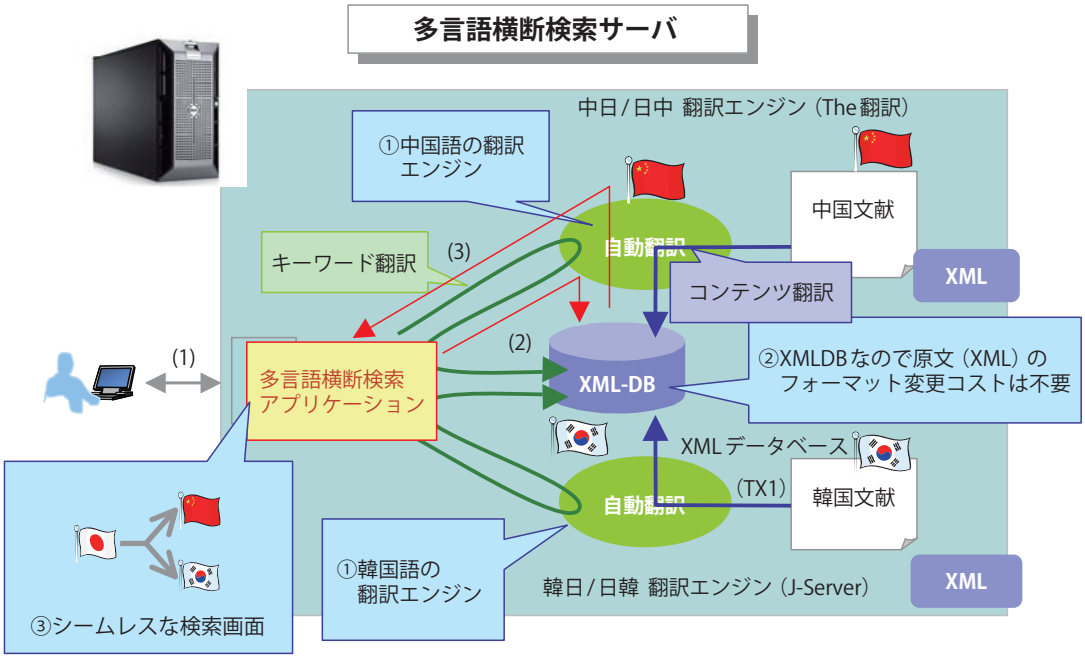

図6　システム概念図

多言語横断検索を行う場合の処理の流れはキーワード翻訳の場合、以下の通りである（図6の中で（ ）数字にて記載）。

（1）アプリケーションサーバを介して検索条件（キーワード／自然文）を入力する。

（2）キーワード／自然文が翻訳エンジンを介して翻訳され、翻訳後の検索条件により、データベースに蓄積された中国文献、韓国文献が検索され、検索後の一覧が表示される。

（3）一覧から出願番号を押下すると当該出願番号の中国文献、韓国文献が日本語に翻訳されて表示される。

6.2 利用イメージ

日本語で中国文献と韓国文献をシームレスに検索する場合の利用イメージを以下に紹介する。

①まず「審査開始画面」にて担当官コードを入力すると「審査状況表示画面」に遷移する。

②「審査状況表示画面」において自然文検索かキーワード検索を選択する。

③いずれの検索方式を選択した場合でも、「検索条件指定画面」（詳細条件指定画面）（図7）に遷移するので、本画面において「IPC」「翻訳方式」「分野別辞書」「検索対象範囲」等を入力し、検索条件を入力後、「検索」ボタンを押下する。ここでは「自然文検索」を選択した後、検索条件として「ラクトフェリンと高分子電解質との複合体」を入力した例を示す。

④キーワード検索を選択した場合は「訳語候補展開」ボタンが押下可能となる。「訳語候補展開」ボタンを押下すると「訳語候補展開画面」（図8）に遷移する。入力したキーワードに対応する訳語が複数表示されるので、検索条件として使用する単語を選択する。ここではキーワードとして、「装置」「言語」「機械」「翻訳」「言葉」「入力」「単語」を入力し、「訳語候補展開」ボタンを押下した場合の画面例を示す。本画面にて検索条件を選択した後、「検索」ボタンを押下する。

⑤いずれの検索方式を選択した場合でも「検索」ボタンを押下すると「検索結果一覧画面」（初期表示画面）（図9）に遷移し、中国文献、韓国文献の検

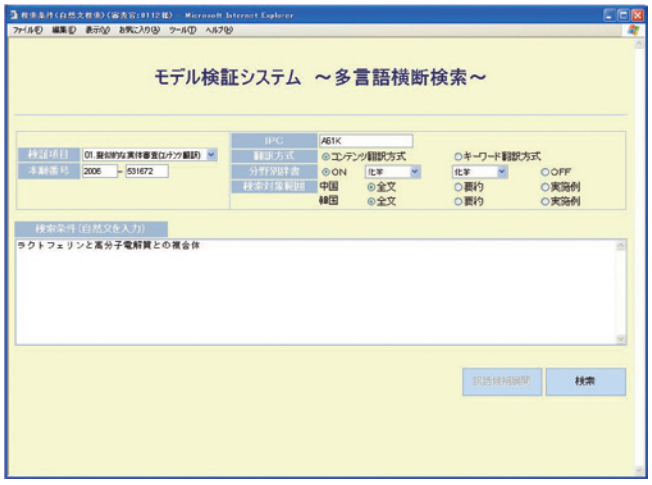

図7　検索条件指定画面

図8　訳語候補展開画面

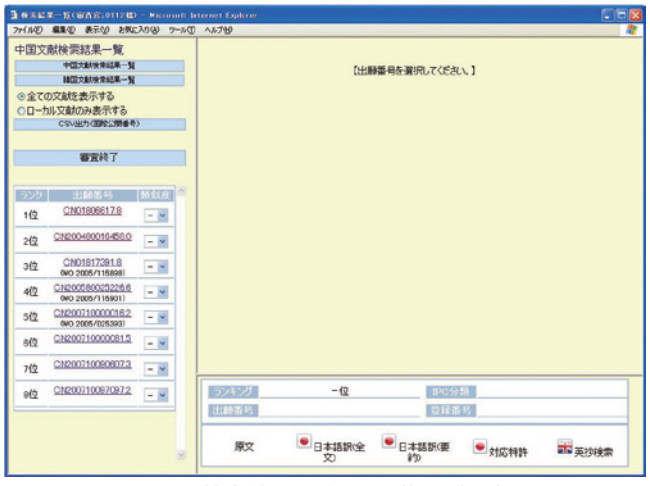

図9　検索結果一覧画面（初期表示）

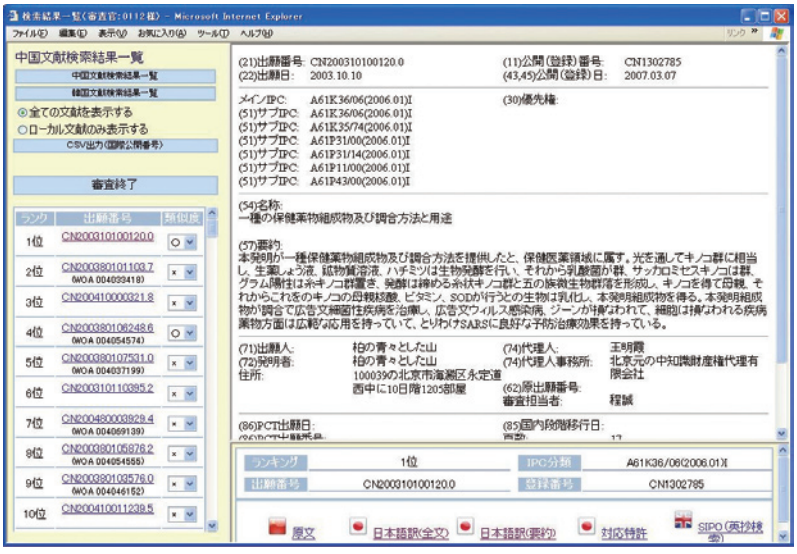

図10　検索結果一覧画面（中国・日本語訳表示）

索結果一覧が表示される。

⑥「検索結果一覧画面」（初期表示画面）において、左下に表示されている出願番号を押下すると当該文献の日本語訳が表示される（図10）。

以上の操作により、中国文献、韓国文献が日本語で検索できるようになる。

7 今後の展望　～長期的な視点～

調査業務を通じて、現状の課題に対する解決策の効果も確認しているが、それでも翻訳技術、検索技術、双方でクリアすべきハードルはまだ存在する。

今回の検証システムでは、以下の対応は実施できていないが、「審査で使える多言語横断検索」を目指すためにも、継続した調査研究が必要と考えている。

7.1 翻訳技術 ～機械翻訳の精度向上～

7.1.1 翻字

(1) 課題～ファミリー文献が無いと対訳抽出できない～

自動辞書メンテナンス機能によって効率的な辞書登録が実現する。しかし、本機能はファミリー文献が存在しない場合、対訳抽出ができず、辞書登録ができない。このため、中国、韓国の国内のみに出願されている分野などでは、的確な効果が得にくい。

(2) 解決策

対訳が存在しないケースへの対応手法として、「翻字（Transliteration）」というアプローチがある。これは、例えば英語の例では、oscarという語が辞書に未登録であっても、英文字のならびから「オスカー」というカタカナ語を生成する処理であり、中国語や韓国語に関しても同様な処理の導入が考えられる[1)]。

しかし、日本語とは発音が異なるため、例えば、ハングルで「데이터」（データ）の各文字を音のままカタカナに置き換えると、「デイト」と誤訳してしまう。長期的には、各言語における発音を考慮した翻字処理の実現／改良が必要となる。

7.2 検索技術 ～情報検索の精度向上～

7.2.1 パラフレーズ検索

(1) 課題～言い回しが異なると正しい検索ができない～

例えば、「デジタルをアナログに変換する」と「アナログをデジタルに変換する」とは違う意味であるが、「デ

1) 中国語では、欧米の人名・地名が音訳される場合がある。例えば、「莫桑比克（ピンイン：mo4sang1bi3ke4）」は、「モザンビーク」の中国語表記で、発音を表す漢字系列になっている。また、韓国語のハングルは表音文字であるので、日本語のカタカナと同様に、外来語を発音のままに表記する場合がある。

ジタルアナログ変換」とはほぼ同じ意味である。このように言い回しが異なる場合、正しく検索できない。

(2) 解決策

この解決策として、日本語では既に研究開発済の機能だが、今回実装しなかった機能（パラフレーズ検索機能）がある。パラフレーズ検索機能とは「デジタルをアナログに変換する」と「アナログをデジタルに変換する」などの言い回しを解析し、区別することにより、検索精度の向上を図る技術である。日本の特許文献を対象としたパラフレーズ検索機能については、その有効性について確認済みである。中国語や韓国語でも一つの意味内容は様々な表現に言い換えて用いられるため、日本語と同様のパラフレーズ検索技術の実現は有効と考える。

7.2.2 各国語に対応した検索エンジンの精度向上

(1) 課題 〜各国語対応は未だ試験的な段階〜

キーワード翻訳型の場合は、今回試験的に中国語、韓国語の対応を実施した。このため、検索精度を上げるべく、外国語に対応した検索エンジンの更なる改良が必要である。

(2) 解決策

文書中に用いられる単語の統計的性質は言語ごとに異なる。このため、検索エンジンの外国語対応として、日本語検索にて培っている最適化パラメータ（単語の出現頻度や文書の長さなどを組み合わせ方の度合い）を言語ごとにチューニングすることで精度向上が望める。

8 おわりに

産業界のグローバル化が進む中、非英語圏も視野に入れた特許審査は、ますます重要性が高くなる。また、ファーストアクション短縮に向けて、現行の審査スピードを落とすことなく、非英語圏の文献のサーチが必要とも考えている。

そして、多言語横断検索技術の両輪である機械翻訳技術、情報検索技術は、現在も日々技術進化を続けている技術であり、数年後には更なる機能向上も期待できる。

このため、日本語で外国文献を検索できる「多言語横断検索技術」は、今後の審査業務に必要なツールになる可能性を秘めており、日本国特許庁の「グローバル化対応力の向上」を推し進めるための強力な武器になると考えている。

弊社も「審査で使える多言語横断検索」システム実現に向けて、今後も継続的な研究開発を進めていく所存である。

profile

住田　一男（すみた　かずお）

1982年　東京工業大学大学院　修士課程修了
1982年　株式会社東芝入社
1999年　東京工業大学大学院　博士（工学）取得
1986年〜2008年
自然言語処理研究に従事

profile

樽井　伸司（たるい　しんじ）

1997年　東芝ソリューション株式会社入社
1997年〜2008年
特許庁審査系システムの設計・開発、インフラ導入等に従事