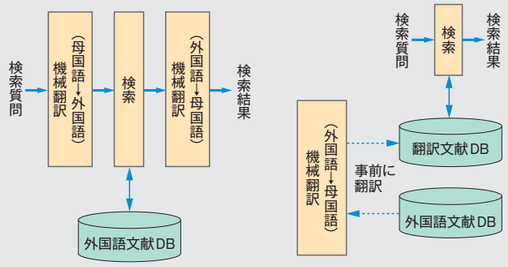
母国語で外国語の情報を収集できる言語横断検索技術

海外の情報に，日本語で簡単にアクセス

インターネットなどによりグローバルな情報へのアクセスが容易になっていますが，必要な情報が外国語でしか存在しない場合や，ワールドワイドな情報収集に迫られる場合が多くあります。しかし，習熟していない外国語の情報を正しく検索することや，かりにそのような情報にアクセスできてもその内容を理解することは困難です。

東芝は，情報アクセスにおけるこのような言語障壁を解消するために，外国語と母国語の情報を区別なく検索できる言語横断検索技術を開発しています。件数が増大している中国や韓国の特許文献の公知例調査で，この技術の有効性が確認できました。



(a) 逐次翻訳方式　　(b) 事前翻訳方式

図1. 言語横断検索の方式 ― 逐次方式では検索質問を検索時に機械翻訳しますが，事前翻訳方式では検索対象の全文書をあらかじめ機械翻訳してDBに登録しておきます。

表1. 逐次翻訳方式と事前翻訳方式の比較

処理方式	長　所	短　所
逐次翻訳	検索時に検索質問の翻訳を行うため，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	検索実行時に翻訳処理も必要なため，検索レスポンスが低下 検索対象の各言語に対応した検索エンジンが必要
事前翻訳	検索エンジンは一つの言語に対応すればよく，システム全体の構成が簡単	新語や翻訳情報の更新に際しDBの再構築が必要 原文表示が必要な場合に翻訳結果とともに原文の蓄積も必要

言語横断検索の方式

近年，経済のグローバル化とコンピュータやネットワークインフラの発達に伴い，各国で作成される技術文書の分量が増加し，加えてそれら技術文書へのアクセスが可能となってきました。しかし，外国語で記載された情報を適切に検索するには，通常，利用者は，キーワードや自然文の検索質問を，その外国語で入力することが必要になります。しかし，その言語に習熟していない利用者にとって，適切な検索質問を入力することは困難です。

言語横断検索技術を用いると，複数の言語で記載された文書を，それらの言語を意識することなく，母国語で検索質問を与えることによって検索が可能になります。更に，母国語に翻訳する機能を用意することで，検索結果の情報を利用者が理解できるように手助けします。この言語横断検索技術は，機械翻訳と情報検索を融合した技術です。

言語横断検索は，翻訳処理の実現形態で次の二つに分類できます(**図1**)。

(1) 検索質問を検索時に機械翻訳する逐次翻訳方式

(2) 検索対象の全文書をあらかじめ機械翻訳して，データベース(DB)に登録しておく事前翻訳方式

これらの方式には長所と短所があり(**表1**)，利用形態を考慮して実現方法を選択する必要があります。

技術課題とアプローチ

●機械翻訳

機械翻訳では，各単語の訳がいかに正しく得られているかが，翻訳及び検索の精度に大きく影響します。特に，特許文献など専門性の強い技術文書では，技術分野ごとに専門用語辞書の整備が重要ですが，そのすべてを人手作業で行うと多大なコストが掛かることになります。

専門用語辞書の登録候補として見出し語とその訳語を対訳文書から自動抽出する，自動辞書メンテナンスを実現することで，このコストの低減を図っています(**図2**)。登録までを完全自動化すると，誤った訳を辞書に登録する危険性があるため，最終的なチェックは人手で行います。

●情報検索

機械翻訳の結果にあいまい性が残ることは避けられません。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などに翻訳できますが，これらの中の一つだけが選ばれて機械翻訳の結果が出力されます。その結果をそのまま利用すると，選ばれなかった訳語の情報は検索されないので，検索漏れが発生する可能性が残ります。

この問題に対処するために，情報検索時に検索語の追加を行うアプローチがあります。検索語追加には，機械翻訳と連携して複数の訳語候補を取得する方法のほかに，疑似適合フィードバックと呼ばれる方法があります。疑似適合フィードバックは，最初に入力された検索質問で初期検索を実行し，その検索結果の上位文書から新しい検索語を選択し，追加します。これらの検索語追加の処理は，機械翻訳の精度が高い英語を対象とした言語横断検索で有効であることを確認しています。また，疑似適合フィードバックでは，検索質問で明示されていない関連語が自動的に追加されるため，言語横断検索に限らずに，平均的な検索精度の向上に効果があります。

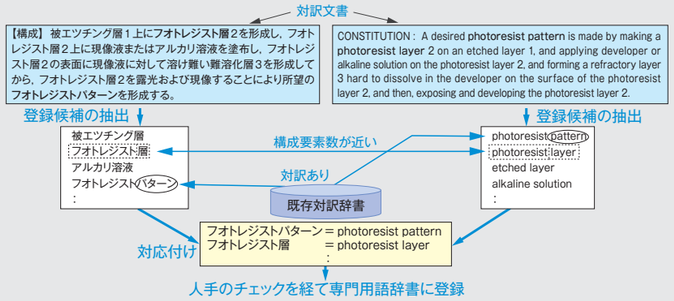


図2. 自動辞書メンテナンスによる専門用語辞書の構築 ― 日本で出願された特許とそれを元に外国出願された特許などの対訳文書から，見出し語とその訳語の登録候補を自動抽出する，自動辞書メンテナンスを実現していますが，最終的なチェックは人手で行っています。

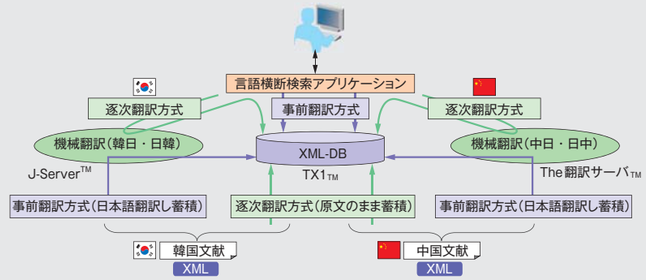


図3. 言語横断検索の実験システム ― 検索エンジンとしてのXML-DBと翻訳エンジンから構成され，中国語と韓国語の特許文献を日本語の請求項やキーワードで横断的に検索できます。

特許審査での有効性調査

近年，中国や韓国での特許出願件数が増大してきており，特許庁の審査業務での公知例調査で，これらの出願特許を調査範囲とする必要性が高まりつつあります。このため，特許庁は2008年度の調査研究として，特許審査業務での言語横断検索の有効性調査を目的とした「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」に取り組みました。

東芝と東芝ソリューション(株)は，特許庁からこの調査研究の委託を受け，中国語及び韓国語の特許文献を検索対象とした言語横断検索の実験システムを開発し，その有効性を調査をしました。このシステムの構築には，中日・日中翻訳エンジンに“The 翻訳サーバ™”のコアロジックを，韓日・日韓翻訳エンジンにJ-Server™(注1)を，検索エンジンとしてXML(Extensible Markup Language)-DB “TX1™”を利用しました(**図3**)。中国語と韓国語の特許文献を，日本語の請求項やキーワードで横断的に検索できます。

キーワードや自然文の検索質問で検索する実験の結果から，自動辞書メンテナンスによる専門用語辞書の構築及び，その中の専門用語を重視した検索処理が精度向上に有効であることが確認できました。しかし，実験期間が限られていたため，特許審査にそのまま適用可能な規模の専門用語辞書は構築できませんでした。専門用語辞書の拡充と整備，及び，その上に立脚した言語横断検索の全体精度の改善は今後の課題です。

今後の展望

産業界のグローバル化が進むなか，ワールドワイドな情報調査の重要性はますます高まると思われます。例えば，今後の特許審査では，現行の審査スピードを落とすことなく，多言語の文献調査を行うことが求められるようになると考えています。

今後，前述の調査研究での知見に基づいて，自動辞書メンテナンスによる専門用語辞書を充実させるなど，様々な技術分野に言語横断検索を適用できるように開発を進めていきます。

真鍋　俊彦

研究開発センター
知識メディアラボラトリー主任研究員

(注1)　J-Serverは，(株)高電社の登録商標。