第 13 回日韓業務交流
業　務　交　流　Ⅱ
2010 年 10 月 14 日

国立国会図書館サーチと日韓自動翻訳の現況及び課題

国立国会図書館　総務部 情報システム課長
中山　正樹

本日は、「国立国会図書館サーチと日韓自動翻訳の現況及び課題」と題して、2010 年 8 月 17 日にプロトタイプ版を試験公開した「国立国会図書館サーチ」（通称：NDL Search）と、その中で実現しようとしている「日中韓翻訳サービス機能」の現状と課題について、お話しします。なお、このシステムは 2012 年 1 月の正式公開を目指して構築中のものであり、今後ユーザ等の評価を受けながら継続的に開発し、機能追加を行っていきます。それらの機能強化の一つとして、日韓翻訳機能の実装を進めています。

1 国立国会図書館サーチ（NDL Search）とは

国立国会図書館サーチは、国内の各機関が持つ豊富な「知」をご活用いただくためのアクセスポイントとなることを目指した検索サービスです。「当館が保有しているか否かを問わず、冊子体に加えて、デジタル化された画像、テキスト、音声等の様々な形態の情報を、いつでも、どこでも、利用者が求める形で、迅速かつ的確に、アクセスまたは案内できるようにすること」を目的としています。

現在、当館が所蔵する図書、雑誌等の資料を探すことができるほか、都道府県立図書館、政令指定都市の市立図書館の蔵書、国立国会図書館デジタルアーカイブポータル（PORTA）が収録している各種のデジタル情報も探すことができます。

2 2012 年 1 月でのサービスイメージ

国立国会図書館サーチは、従来の図書館利用者、図書館員の方だけではなく、広く一般の利用者、各種の Web サービスを提供している個人、企業・団体の方などの利用者も含め、幅広い範囲の方々に利用していただくサービスの提供を目指しています。国立国会図書館サーチの用途として、以下のようなケースを想定しています。

- 一般的な検索・閲覧（GUI での利用）
  - 当館が所蔵する各種資料を対象とした検索を Web 上で行う（NDL-OPAC）
  - 当館が所蔵する雑誌記事を対象とした検索を Web 上で行う（雑誌記事索引）
  - 日本中の図書館が所蔵する図書を対象とした検索を Web 上で行う（公共図書館総合目録）
  - 当館および提携機関が所蔵するデジタル資料やレファレンス記録などを、図書や雑誌記事と併せて Web 上で検索する（統合検索サービス）
- 情報の再利用（API での利用）

- 国立国会図書館サーチの検索結果を、利用者自身のWebサービス上で利用する（検索結果API提供機能）
- 国立国会図書館サーチの収録データを、まとめて入手して利用する（メタデータダウンロード機能）

- サービス・機能の再利用
  - 国立国会図書館サーチで開発されたシステムを使って、各機関・企業が作成したデータと併せて提供する（マッシュアップによるサービス提供支援）
  - 国立国会図書館サーチのシステムとデータを研究・開発に利用する（テストベッド環境の提供）
  - 国立国会図書館サーチのために開発されたオープンソース・ソフトウェアを利用する（図書館システムとしてオープンソフトウェア（OSS）で提供）

3 2009年度開発版の到達点

2009年度開発版システムを2010年8月に試用公開しました。公開当初から予想外に多くのアクセスがあり、レスポンスの低下を招きましたが、各種マスメディア、Twitter、ブログ、ソーシャルブックマーク等でも多く言及され、概ね好意的な評価をいただきました。検索システムとして、基本的な機能は実装されており、また、書誌情報の定期的な登録更新も行っていますので、レスポンス等が改善されれば、十分実利用に耐えうるものと認識しています。提供している主な機能とシステムのイメージは次の通りです。

(1) 主なサービス

- 基本機能
  - 現在、34個のデータベースから収集した約5,500万件の文献情報等を検索できます。（検索対象は、http://iss.ndl.go.jp/information/target/ を参照）
  - 全文テキスト化された資料に関しては、書誌情報だけでなく、本文の全文検索ができます。
  - 統合検索の結果について、可能な限り入手手段を案内します。（近くの図書館、Amazon、GoogleBookSearch等へもナビゲート）
- 検索支援機能
  - 連想検索、類義語・同義語検索等を用いて検索を支援します。（あいまい検索機能）
  - 「日本語 ⇔ 中国語」「日本語 ⇔ 英語」の翻訳検索・翻訳表示ができます。（翻訳機能）
- 検索結果のグルーピング機能
  - 複数の機関で所蔵している同一の資料をまとめて表示します。（書誌同定機能）
  - 形態を異にする同一著作を隣接表示します。（著作単位でのグルーピング）
  - 検索結果は、適合度順（検索語に対する各資料の関連性が高いもの順）で排列します。
- 絞り込み機能と再検索機能
  - 資料種別、所蔵館等から絞り込み検索を行うことができます。（ファセット検索）
  - 関連キーワード等から再検索を行うことができます。（シソーラス検索、連想検索等）
- ブックマーク機能
  - 書誌情報の固定URL表示、Twitterへの投稿機能、検索結果一覧の動的RSSの配信機能等、検索結果を活用するための様々な付帯機能を利用できます。
- 外部サービス連携機能
  - このサービスを他のシステムから利用するための各種標準的なAPIが利用できます。

(2) システムイメージ

国立国会図書館サーチ（開発版）では、オープンソースの統合図書館システムであるNext-L Enju[1]をコアに、Heritrix、Hadoop、GETAssoc[2]、WordPressといったOSSを活用してシステムを構築しています。開発した部分のソフトウェアは、将来的にはOSSとして公開し、公共図書館等での利用に供することを想定しています。（別紙システム構成図を参照）

4 2010 年度、2011 年度の開発予定

試用公開したシステムについては、ユーザの意見を反映し、より使いやすいシステムに改善すると共に、今後、インターネットで普及しているサービス、実証実験等で実用化が検証された技術を積極的に採用し、順次、機能強化を行う予定です。現在想定している主な機能強化項目は以下の通りです。

(1) 2010 年度機能拡張開発

検索、収集・組織化、ナビゲーション、利用者付加価値機能等、本システムの中核となる利用者共通的な機能、性能改善。

(2) 2011 年度機能拡張開発

国会関係者・児童・障害者・来館者等、独自のインターフェースを必要とする利用者のための機能。

5 外部機関・サービスとの連携方針

デジタルネットワーク時代に、利用者に求められるサービスと機能を持ったシステムを構築し提供するためには、外部の機関との連携協力が必須であり、当館は積極的に連携協力を行っていきます。その連携の姿勢として、次のような方針を掲げています。

(1) メタデータの収集または横断検索等による統合検索サービスの提供

外部機関・サービスが提供するコンテンツのメタデータを当該機関・サービスの許諾を得て収集、もしくは横断検索します。

(2) 外部のウェブサービスとの連携によるサービスの提供（マッシュアップサービス）

外部で提供されている連想検索サービスや機械翻訳サービス等のウェブサービスを有機的に組み合わせて、付加価値の高い検索サービスを実現します。また、外部の情報サービスへの効果的なナビゲーションを実現することにより、利用者の情報探索を支援します。

(3) コンテンツの統合利用促進のための環境整備

有用なコンテンツを保有しているにもかかわらず、データベースの構築や検索サービスの提供ができない機関に対して、データベースの構築や API 実装等を支援します。

(4) 研究開発、技術開発における連携

利便性の高いシステム構築のためには、現状で確立した技術のみでは実現が困難です。大学の研究室、官民の研究機関、ベンチャー企業等による各種の情報技術に係る研究開発を支援するために、当館の情報資源を利用した実用化・実証実験を行うことができるよう、テストベッドの場を提供します。

---

[1] Next-L Enju とは、大学等の研究者が開発したオープンソースの図書館システム。

[2] GETAssoc とは、国立情報学研究所（NII）連想情報学研究開発センターが開発した連想検索エンジン。

6 実施中、実施予定の実証実験

実用化実証実験として、現在実施中、実施予定のものは以下の通りです。

- 県域の市町村立図書館蔵書目録の検索
- 書誌同定・集約表示の精度向上
- シングルサインオンの実現方式の検証
- 全文テキスト化・全文検索
- 日中韓翻訳機能

7 日中韓翻訳実験

(1) 経緯

NDL Search の検索対象の多くは日本語の資料ですが、当館のアジア言語 OPAC をはじめ、統合検索対象には、韓国語、中国語の文献も含まれます。利用者が日本語で検索しても、他言語の資料は検索・閲覧できず、その言語で検索しなければなりません。そこで、日本語で入力しても他の言語に翻訳して検索し、その検索結果も日本語に翻訳できれば、他の言語の資料も利用が容易になると考えました。

(2) 2009 年度システムでの実装

2009 年度開発システムにおいては、情報通信研究機構(NICT)の「日中・中日言語処理技術の開発研究」(http://www2.nict.go.jp/x/x161/project.html)の実証実験システムを、WebAPI 経由で利用させていただき、検索キーワードを日本語から中国語に変換して、中国語文献等の検索を行えるようにしました。この実証実験システムは、日中の汎用的な用例翻訳手法および科学技術文献の翻訳・情報検索用辞書を半自動作成する手法を利用したものです。また、現在の NICT の実証実験システムでは、科学技術分野以外の自然文翻訳は困難なため、書誌詳細画面での翻訳は、Google の翻訳サービスを利用して、書誌項目を中国語から日本語に翻訳できるようにしました。

(3) 2010 年度システムでの実装

2009 年度のシステムでは、中国語翻訳機能しか実装できませんでしたが、日中韓の言語の壁を越えることは重要なことであり、韓国語に関しても同様に進めたいと考え、後になってしまいましたが、韓国語翻訳の実装も進めています。

NDL Search では、今後の拡張性として、複数の翻訳サービスを自由に選択できるように開発しており、現在、韓国語も翻訳できる翻訳サービスとして実績のある(株)高電社の翻訳ソフトウェア(http://www.kodensha.jp/)を採用することとして、追加実装を行っています。

(株)高電社は、1979 年に設立され、1990 年より日韓翻訳システムの開発に着手しており、現在、韓国語、中国語、英語と日本語の双方向のテキスト翻訳機能(単語及び自然文翻訳)及びウェブページ翻訳機能(ウェブページ内全文テキスト翻訳)を、パッケージソフトウェア及びウェブサービス(ASP)の形で提供しています。構文解析においては、ルールベースを利用して、助詞、助動詞に関わる手法など韓国語の特性に合わせた言語処理と、意味解析辞書による訳語の最適化処理が特徴となっています。基本辞書としては、韓国語は、日韓 35 万語、韓日 27 万語、中国語は、日中 35 万語、中日 29 万語を持ち、ユーザ辞書も追加可能になっており、日中翻訳で 80%以上、日韓翻訳で 90%以上の認識率を持つということです。

地方自治体の日本語ホームページを英語・韓国語・中国語に自動的に翻訳したり、大手のポータルサイトへの翻訳サーバの導入実績があると聞いています。

これにより、韓国語や中国語の文献を日本語で検索し、検索結果を日本語に翻訳することも可能になります。また、日本語の文献を、韓国語や中国語で検索することも可能になります。

(4) 翻訳サービスの流れ(別紙 PPT)

翻訳サービスの流れのイメージは別紙のとおりです。

NDL Search の検索画面、書誌詳細画面、情報保有サイト(目録、本文、ウェブページ)(実装準備中)において、翻訳機能を実装しています。①検索画面においてキーワードを入力して、「日本語→韓国語」「日本語→中国語」、を選択して、検索実行をします。②キーワードが、テキスト翻訳サーバにより、翻訳されて、③その翻訳されたキーワードにより、文献 DB を横断検索します。④検索先から検索結果の書誌情報(原文)を得て、利用者画面に表示されます。ここで、⑤「翻訳ボタン」を押下することにより、書誌情報(原文)がテキスト翻訳サーバで翻訳され、利用者画面に書誌情報(翻訳結果)が表示されます。さらに、書誌情報(翻訳結果)にある文献ページへの「翻訳リンクボタン」を押下することにより、そのページの全文テキストを翻訳して表示されます。

(5) 日中韓三か国による検索、翻訳の実験(提案)

8 月に締結された「日中韓電子図書館イニシアチブ協定」で掲げられた目標の実現策として、

- NLC、NDL 及び NLK は、メタデータ・スキーマの標準化、情報サービス(ポータル)の統合及び電子情報への長期のアクセスを可能にするための共同開発を促進する。
- 各図書館が有するそれぞれのポータルの相互運用性を高めることにより、三者の統合的情報サービスの第一段階を実現する。

と掲げられましたが、この合意に基づく最初の共同実験として、日本語、中国語、韓国語の言語の壁を超えた実験を提案させていただいています。

NDL Search では、現在、日本国内の統合検索先が保有している韓国語、中国語の資料の翻訳検索を行なえるようにしようとしているところですが、今後、双方向での検索・閲覧を実現することを視野に、まずは、翻訳機能の実用性の検証のために、NLK、NLC の蔵書目録等を検索・閲覧対象としてさせていただきたいと考え、下記のような提案をさせていただいております。

- NDL Search から、NLK、NLC の蔵書目録を横断検索させていただきたい。
- そのために、利用可能な通信プロトコル(SRU/SRW または Z39.50 等)の種別と、アクセスのための URL 等の情報を教えていただきたい。

6 月から、NLC、NLK に対して、それぞれ担当者レベルで具体的な調整をさせていただいています。

次のステップとしては、韓国、中国で実験的に利用可能な翻訳システムがあれば同様に実装させていただくこと、また、翻訳精度をより高めるために、三か国協力して、用例集、用語の語彙の充実に取り組みたいと思っています。

この成果を各国のポータルシステムに実装することにより、言語の壁を越えて、各国が保有している情報を相互に検索・閲覧できることを目指したいと思います。

8 ポータルの相互運用性を高めるために

NDL Search は、まずは、国内の様々な機関が保有する情報資源を、「いつでも、どこでも、誰でも、所蔵機関や媒体の形態を問わず、「情報」を閲覧もしくはナビゲーションする」ことを目指しています。そこに、翻訳機能が実用化されれば、「言語を問わず」も目指すことができます。

今後はさらに、各図書館の資料が同一もしくは類似の資料と判断できるように、交換されるメタデータの記述要素、記述規則の共通化や、類義語を把握するシソーラス、語彙の違いを吸収するオントロジー等の言語や表現の差異を吸収する技術を相互交換して適用することにより、より的確な検索が可能になるようにしたいと考えています。これにより、三者の統合的情報サービスの第一段階の実現に近づけるものと考えています。

9 情報の利活用の促進を目指して（課題と今後）

最後に、利用者の情報探索の目的は、問題・課題の解決であり、回答が掲載された資料の所在ではなく、回答そのものを知識として得ることです。現在では、多くの利用者がインターネットで閲覧できる情報だけで問題を解決しつつあり、Google などの検索エンジンで見えない情報はないも同然と言われています。こうした情報環境の中で、知識・情報の利活用を促進するためには、紙の形態で流通している資料もデジタル化し、内容をインターネット上で検索・閲覧できるようにする（可視化）ことが必要です。

今後の課題として、従来の単なる情報探索から、事実としての知識検索へ進化させ、知識の再利用による新たな知識の創造が求められています。それを実現するためには、単に資料の内容を可視化して集積するだけでなく、個別の情報に意味的にタグ付け（自動組織化）し、知識として相互に関連付けて（自動集合知化）、利用者が求める知識として、より的確に取り出せるようにすること、また、知識として有効に活用するために、情報の信頼性を確保することが必要です。

「可視化」において翻訳により言語差異を吸収し、また、「組織化」、「集合知化」等において情報の記述要素、記述規則等の差異を吸収することができれば、「国の文化・科学遺産への容易なインターネット・アクセスを提供すること、豊かな多文化・多言語コンテンツを一般の人々に享受してもらうこと、学術世界へ貢献することを目的とする。」という日中韓電子図書館イニシアチブ協定の目標の達成の一翼を担えると考えています。