2007－CE－91 (6)
2007／10／5

多言語利用可能なオントロジーを使った教育コンテンツ検索システムの開発と評価

千葉栄一、野殿浩一、上田幸宏（NTT コミュニケーションズ）
川森雅仁（NTT サイバーソリューション研究所）、亀山渉（早稲田大学）
古屋和行、原山唱一（NTT ソフトウェア）

アジア地域の教育レベルの情報格差を是正する手段として、遠隔教育の需要が高まっているが、言語、文化等の差異が遠隔教育普及への大きな課題となっている。そこで、言語、文化等の差異を意識することなく教育コンテンツを検索できるプラットフォームを開発し、その有効性について検証・評価を行った。

Development and the evaluation of the education contents search system which uses the multilingual ontology

Eiichi Chiba, Kouichi Nodono, Yukihiro Ueda (NTT Communications)
Masahito Kawamori(NTT Cybersolutions laboratory), Wataru Kameyama (Waseda University),Kazuyuki Furuya, Shouichi Harayama (NTT Software)

As means to minimize gaps of educational information/contents among the Asian regions, a demand for long-distance education is rising. However, language and culture differences cause serious problems to apply the education style widely in Asian countries. Therefore we developed the platform for searching education contents without being conscious of language and culture differences and evaluated its effectiveness.

1.はじめに

現状、アジア各国間には顕著な情報格差があり、この格差はアジア域内の産業や文化を活性化するうえでの大きな障壁となっている。情報格差が発生する要因として、文化の違い（≒表現の違い）、言語の違い等によって、アジア域内に情報が存在するにも関わらず、必要となる教育コンテンツ等が見つけられず、取り出せない環境が課題として挙げられる。更に近年では動画像等を、教育素材として採用している教育機関もあり、教育コンテンツの多国間での共有利用の環境も望まれつつある。

これら課題の解決に向け、文化・言語の違いを意識することなく、それぞれの母国語である日本語、タイ語で教育コンテンツを検索できる、“オントロジーベース検索機能”を開発し、日本語、タイ語でのネットワークを介した検索において、文化、言語の違いの吸収を実現した（図 1）。本文ではその実験結果・考察を含め、実験概要について述べる。

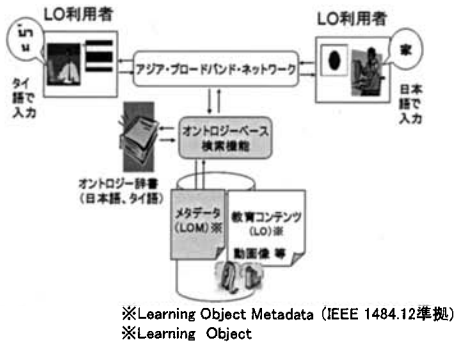

図１　実証実験全体イメージ

2.プラットフォーム構成要素

2.1　オントロジーとは何か

オントロジーとは、哲学の一分野としての存在論として「機械に可能な、用語や概念間の関係を明確に定義した規定（あるいはそれを実装したシステム)」という意味である。本文でも Web 等で知識の共有を実現するために、あるデータ A とデータ B との論理的な関係を記述するための枠組みとして捉えており、オントロジー記述言語として W3C 勧告でセマンティック Web 構想の一部である、OWL を XML 中に定義した。(図 2)

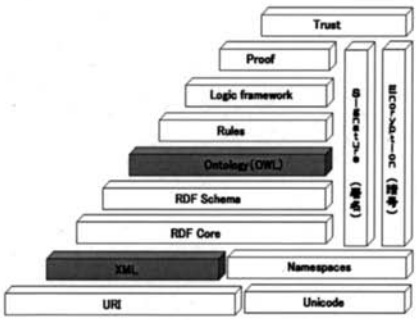

図2　セマンティックWebの階層

2.2　オントロジーベース検索機能

オントロジーベース検索機能では、それぞれ日本語、タイ語からの入力キーワードから、オントロジー変換を行いメタデータ（LOM）からの検索を可能とするために、日本語、タイ語のオントロジー辞書を作成した（図 3)。　日本、タイ両国における、文化・言語の違いを吸収する手法として、文化の違いの吸収ではオントロジー辞書の拡張・マルチドメイン化、言語の違いの吸収では、概念の ID 化による多言語対応を行い、それを実証する為の検証項目を導出した（図 4)。

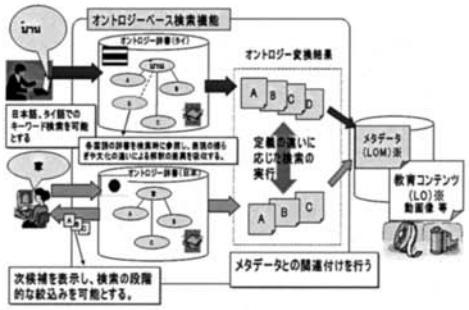

図3　オントロジーベース検索機能

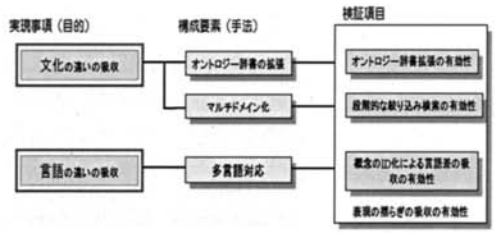

図４　構成要素（手法）と検証項目

2.3　メタデータ

本実証実験では、オントロジー変換結果と教育コンテンツ（LO：Learning Object）を結びつける要素として、メタデータ（LOM：Learning Object　Metadata）を使用している。この LOM は IEEE1484.12 として標準化されており、各教育機関での LO の共有利用を促進させることを想定し採用し

た。

また、本実証実験での検索対象となるLOのドメイン（分野ごとに体系化された概念の集合）は、日本・タイ両国の文化的背景による特徴(類似点や相違点)が顕著であり、異国間相互理解を題材としたコンテンツとして、“食（料理)”をターゲットドメインとした。そして食（料理）に関連のあるドメインを中心にオントロジー辞書のマルチドメイン化を行い、その結果、「食材」「道具」「医学」等18ドメインとなり、ドメインに関するオントロジー辞書の定義数は、日本・タイ合計で1600個となった。

3.遠隔教育コンテンツ高度検索技術の実証実験

3.1　システム構成

実証実験では、オントロジーベース検索機能・LO/LOMを格納するサーバを日本側データセンターに設置し（図5)、教育機関からのキーワード検索として、日本側から早稲田大学、タイ側からはチュラロンコン大学の支援のもと、図4の4点の検証項目について結論を導出できるよう、各パターン別に各々20題の課題を設定し実験を行った。

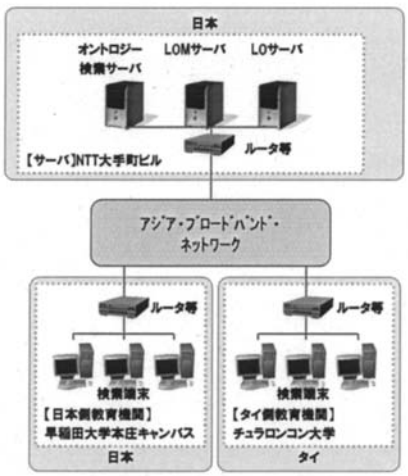

図5　実証実験システム構成図

3.2　構成要素技術

3.2.1 文化の違いの吸収

文化の違いの吸収を行うためオントロジー辞書の拡張、段階的な絞込み検索を検証項目として実験を行った。以下にそれぞれの構成要素技術を示す。

(1)オントロジー辞書拡張

オントロジー辞書は、その定義が疎（少ない）の場合、オントロジーベース検索の効果は限定的になると考えられる。しかし、表現の揺らぎや概念間の一般的な関連だけでなく、ある文化において自然な発想、ある文化固有の関連性を定義した密なオントロジー辞書であれば、その文化が持つ概念体系を実現し、利用者にとって使いやすい検索機能を提供できる。(図6)　本実験での“表現の揺らぎ”とは、表記の揺れ（漢字、ひらがな、カナでの複数の表記方法)、同義語（発音は異なるが、同じ意味を持つ言葉であり、外来語、省略形、通称が該当)、コード表現の揺れ（コンピュータ上の文字コード）を指し、基礎となるオントロジー辞書の作成時に定義、または、オントロジー辞書の拡張時にも追加定義を行った。

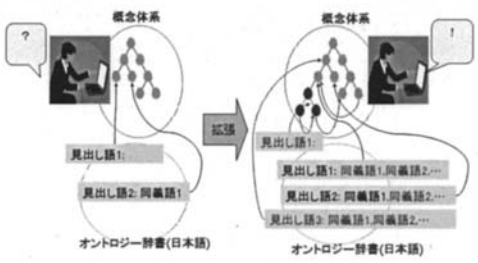

図6　オントロジー辞書拡張イメージ

(2)段階的な絞込み検索

本実験では、“食（料理）”をターゲットドメインとしマルチドメイン化したが、マルチドメイン化とは複数のドメインを横断した検索が可能なことを指し、文化の中でも分野の違いの吸収に該当する。LO利用者が求める有益な情報にたどり着ける仕組みとして、段階的な検索インターフェースを作成し、オントロジー辞書に基づき、検索キーワードに関連する語句や概念の候補を表示し、情報へのナビゲーションを行った。

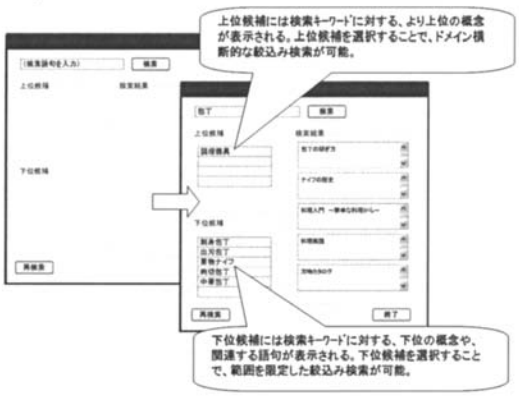

図7　検索画面イメージ

3.2.2　言語の違いの吸収

言語の違いの吸収を行うために、言語ごとにオントロジー辞書を保持することで、各母国語（日本語、タイ語）での検索を想定した仕組みを構築し、LOの他国間共有を実現した。オントロジー辞書は、言語に依存しない形式（概念ID）を用いることで概念として体系化する。これにより各母国語での検索に応じて、語句自体だけでなく、該当する概念 ID をオントロジー辞書から抽出し、その概念に結びついた他国のLOを検索可能とした（図8）。一例として、日本語検索した場合、日本語オントロジー辞書では概念ID：333、555がオントロジー変換結果として抽出され、そのID：333に該当する日本、タイのLOMが検索される仕組みである。

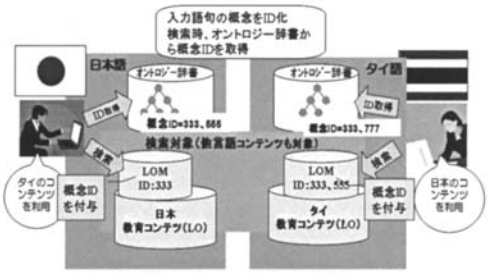

図8　概念のID化による多言語対応イメージ

3.2.3　実験結果と考察

実験結果詳細は別途報告するが、ここでは検証項目別の考察と主な課題について示す。

①　表現の揺らぎの吸収では、オントロジー辞書に表現の揺らぎを定義することで、検索キーワードの入力インターフェースへの有効性を確認できた。全体傾向としては、“表記の揺れ”が最も多く発生していた。表1は、日本、タイ両国での課題別の表現の揺らぎの吸収割合を示したものである。J07 課題では、延べ41回の検索キーワードが入力され、そのうちの約42%にあたる17回の表現の揺らぎ（例：「日本料理」と「和食」）が発生し、約32%にあたる13回はオントロジー辞書により表現の揺らぎが吸収されていた。

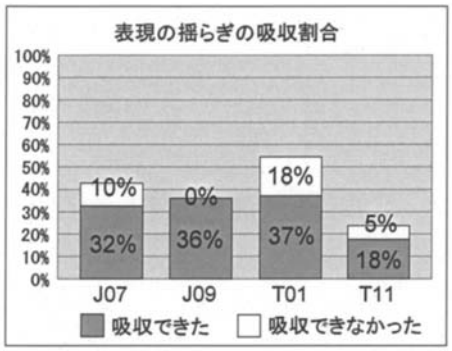

図 9　表現の揺らぎの吸収割合

②　日本、タイの LO 利用者が入力した検索キーワードや検索経路を分析したうえで、オントロジー辞書に対する概念や概念間の関係の拡張が、検索効率の向上に大きく効果があることが確認できた。図 10 は、日本、タイ両国で平均何回目の検索で、「適合」した LO を見つけることができたのか、オントロジー辞書拡張の前後で比較を行った。値が小さいほど良好な結果と言える。これらから、オントロジー辞書に各文化のオントロジーを反映し、且つこれを成長させていく仕組みの実現が、文化の差異の吸収には重要な要素であるとの認識に至った。

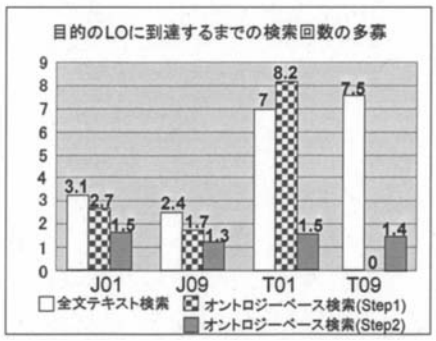

図 10　目的の LO に到達するまでの検索回数の多寡

③　段階的な絞込み検索では、概念同士の上位・下位の関係、ドメイン間の横の関係を定義し段階的な絞込みを行うことで、検索結果に対する適合割合が高くなり、有効であることが確認できた。表 2 は段階的な絞込み検索有無別の、検索結果の適合数を示しており、適合する LO を見つけられた検索回数の割合は、絞込みを行った場合と行わない場合でそれぞれ約 91%、60%の結果となった。一方、検索効率の観点として概念同士の横の関係から、分野を跨った LO 数増大時に、段階的な絞込みだけでは十分に絞込めず、逆に利便性が低下する可能性も否定できない。

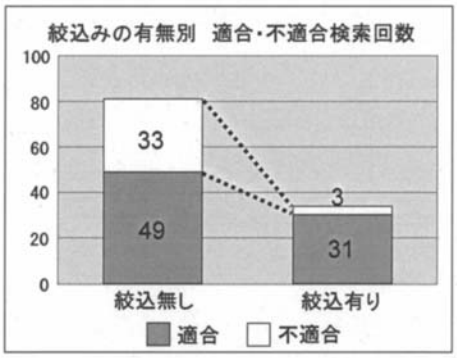

図 11　絞込み有無別の適合・不適合検索回数

④　異国間相互理解のため、文化の差が出やすい“食（料理）”のドメインを中心にマルチドメイン化をし、日本語・タイ語のどちらの言語で検索しても、互いに相手国の LO が検索できることを確認した。ここでは言語の差異吸収の構成要素として、概念の ID 化による多言語対応の仕組みが機能したと言

える。しかしながら、概念のID化、及びその体系化に相応の手間（コスト）が必要になることも課題であり、概念のID化をいかに手間をかけず効率化するか、その手法の確立について更なる検証が必要である。

4.まとめ

本実証実験では、遠隔教育の場における教育コンテンツ（LO）のオントロジーベース検索機能の有効性については立証できた。ただ、実用レベルまでには検証すべき課題もあり、特に利用者定義によるオントロジー辞書の自律成長をさせ、教育コンテンツ（LO）の検索に生かす仕組みは、運用面を考慮したオントロジーベース検索機能を継続使用するために検証すべき要素であり、以降の取組みとして検証していきたい。

謝辞：本文は、平成18年度の総務省プロジェクト「国際情報通信ハブ形成のための高度IT共同実験」の一環として実施されたものである。

参考文献：デジタル・コンテンツ流通教科書　亀山渉監修