次世代インターフェースとしての多言語コンシェルジュの研究開発（071703017）

Multilingual Concierge as Next Generation Human Interface

研究代表者
町田和彦　東京外国語大学
Kazuhiko MACHIDA　Tokyo University of Foreign Studies

研究分担者
ペーリ・バースカララオ†　高島淳††
Peri BHASKARAROA†　Jun TAKASHIMA††
†東京外国語大学　††東京外国語大学
†Tokyo University of Foreign Studies　††Tokyo University of Foreign Studies

研究期間　平成 19 年度～平成 21 年度

概要

「多言語コンシェルジュ」とは、多言語・多文字で表現された情報資源に容易にアクセスするためのサービスとそのヒューマンインターフェースの総称である。本研究では、多言語・多文字環境でのデータ蓄積および Unicode 運用上の問題点を抽出し、文字コード列生成方法の仕様策定を行う。これらの成果の上に、対象言語・文字に不案内なユーザでも、あいまいなつづり、自身の母語、既知の外国語等の言語情報を手がかりとして、言語や文字の壁を意識しないで多言語・多文字情報資源へのアクセスを可能にするサーバー/クライアントモデルの Input Method を提供する。

Abstract

Our concept, ‘Multilingual Concierge’ covers the whole system of services combined with human interfaces to provide easy access to information resources expressed in various scripts appropriate to the languages. The present study aims, first, at investigating crucial issues underlying multilingual data storage in Unicode encoding, and at fixing methods to generate allowed code sequences. On the basis of investigations and experiments, the objective of our study is to provide a practical and multipurpose IME (Input Method Editor) which is based on Server/Client model. This IME enables users, who do not have sufficient knowledge about target languages and scripts, to access multilingual information resources with the help of the knowledge of either mother tongue or foreign languages they know more or less, as well as with the clue of ambiguous spellings inputted.

１．まえがき

「多言語コンシェルジュ」は不特定の言語・文字で表現された情報資源を言語や文字の壁を意識することなしに利用できる環境を提供し、言語バリアを取り除く近未来の中核的システムと位置付けられる。グローバリズムの波が押し寄せる現在、特に多言語・多文字環境における文字情報への対応が急務である。

モデルシステムの実証と実装を目的とする本研究では、世界の主要文字系統であるラテン文字（キリル文字を含む）、インド系文字、アラビア系文字、漢字の 4 大文字圏における 24 言語を実証実験の対象とした。

２．研究内容及び成果

当初の目標・計画に従って、初年度に本研究に必須なテストベッドサーバーを構築し、以降順次ソフトウエアおよび機械辞書の開発・実装に努めた。

本研究は 2 つのコンポーネント、「1) 言語コンポーネント」と「2) 多文字環境に適した Input Method」で構成されている。前者は多言語コンシェルジュを支える文字コード列処理方法の調査・検討・解決方針に関係し、後者は最終目標である実用的なシステムそのものの構築と関係している。

多言語コンシェルジュでは、図 1 で示すように、 検索プロセス全体における文字コード列を 3 つの異なる立場から考えている。アクセス（検索）するユーザが IME を利用して入力する文字コード列、それを受け取り検索エンジンが検索する文字コード列、検索対象である情報資源に含まれる実際の文字コード列である。

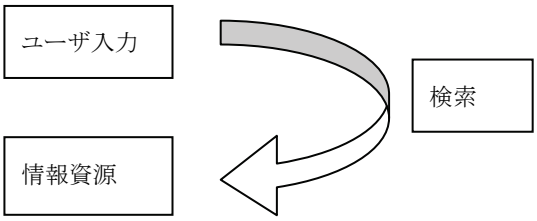


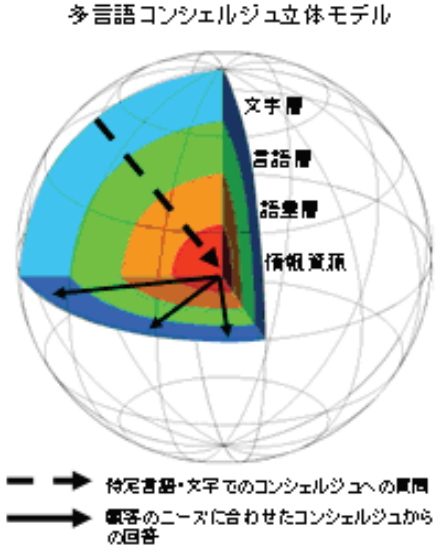
図 1 検索プロセスにおける文字コード列

ユーザが期待する検索結果を得るためには、入力・検索・情報資源それぞれの局面において、文字コード列の正確さが前提とされると同時にある程度の柔軟さが要求される。この問題には、自然言語の慣習（例えば表記のゆれ）と文字コードの標準化という性質の異なる問題が同時に含まれていることに着眼し、言語・文字体系ごとに問題のレベルの設定をすることで解決を試みた。

「1) 言語コンポーネント」では、特に言語固有の「つづり」と密接な関係にある「言語・文字処理系の中で自動的な標準化」と、文字コードに内在する「Unicode 正規化」の問題を扱った。具体的には、現在の web 上に公開されている多言語・多文字情報資源と既成の検索エンジンとの組み合わせによる問題の質と量の調査研究を進めながら、現実的な問題解決の方向を探った。公開されている情報資源には、程度の差はあれ、多種多様の問題が潜んでいることが確認された。この問題はユーザ入力と同時に、情報資

源構築（＝入力）に使用されている現行の不特定多数の IME に起因することも確認された。これは、ユーザの不注意や意図的な悪意の有無は別にして、セキュリティー上危険な「なりすまし」(spoofing)の原因にもなる。また複数の既成の検索エンジンを、いろいろな条件下で使用した検索結果を基に検証したところ、Unicode 正規化を含めた性能実装が一様ではないことも確認した。

「2)多文字環境に適した Input Method」の中心コンセプトとして、メタ・スクリプト(metascript)と呼ぶ新しい概念を導入した。これ自体は、Unicode がサポートするすべての文字入力がキーボードから直接入力できる 7bit の ASCII 文字（ラテン文字）で構成される記号セットである。メタ・スクリプトをいわゆる転写(transliteration)と区別するのは、この概念が転写文字のような規範性の強い固定したものではなく、文字あるいは語との関係（割り当て）において、目的に応じて柔軟に設定可能な仮想文字列という性質をもっているからである。例えば表音文字では特定の音特徴（母音、子音、アクセント、トーン、クラスターなど）にメタ・スクリプトを設定することも可能であり、また「逆引き」、シソーラス、多言語対訳のように使用する辞書によってメタ・スクリプトの概念を語のレベルにまで拡張することもできる。



多言語コンシェルジュための Input Method 開発は、本研究の研究成果の集大成であり、また実用化モデルの実装でもある。サーバ/クライアント型のモデル実装には動的に HTML を生成するサーバサイド・スクリプト言語として PHP(Hypertext Preprocessor)を採用した。このシステムは機能別に集約したツール群である (A)多言語・多文字入力装置、(B)全文検索フェーズ、(C)言語解析ツールから構成される。これらのツールは単体としても利用できるが、もともとは入力から検索・解析へと連動する一貫した処理系として設計されている。

(A)入力機能(IME)には、メタ・スクリプトを応用した多言語・多文字入力の実証実験をし、十分実用レベルにあることを確認した。また Unicode 正規化を実装した入力文字チェック機能、レーベンシュタイン距離を応用した綴り字チェック機能も開発した。(B)検索機能は、多言語・多文字環境での通常文字列検索のほか、正規表現検索、シソーラス検索(多言語対訳、類語、綴りのゆれなどに対応)、頻度の高い語との共起確率条件を優先する検索などが選択できる。本研究の発展拡張として開発した(C)言語解析ツールは、N グラム方式の言語判定を補完する機能と、一部の言語では実用的なレベルの語形・句構造解析機能を備えている。

本研究期間中、実証研究のために開発・構築した多言語・多文字機械辞書の数は 100 以上にのぼる。機械辞書の構造は、将来の修正・追加・コンバートが容易であることを考慮してシンプルになっている。

３．むすび

グローバリゼーションの波の中で、言語と文字のバリアを越えた言語情報資源へのアクセス（検索）サービスの中核技術である多言語コンシェルジュというコンセプトはこれからますます重要になると考えられる。言語学では言語＝音声という関係が前提となっているが、21 世紀のネットワーク環境では入力・検索・情報資源すべての局面において言語＝文字という人類史上未曾有な関係が常態化しつつある。こうした背景の中で本研究が目標とした文字（正確には文字コード）中心に据えた処理系は、ネットワークを利用するすべての分野と関係するといっても過言ではない。

開発された入力に関する技術(IME)は多言語・多文字入力の新規サービスの可能性に直結している。開発したシステムは利便性だけでなく、現在事実上の標準である文字コード Unicode に内在する望ましくない曖昧性と危険性を極力排除できることも検証できた。これにより、ユーザの検索文字列の多言語・多文字入力だけでなく、セキュリティーの高さが要求されるポータルサイトの情報資源構築にも適している。

開発された検索機能は、文字通り言語と文字のバリアを超えた検索を可能にする。電子図書館を例にとれば、ユーザは自分が得意とする言語・文字であらゆる分野のあらゆる言語・文字で書かれた文献にアクセスできることになる。特に本研究でシソーラスと呼ぶオプション検索辞書の概念は、ニーズに合った各種辞書を実装することで多言語対訳、類義語、反対語、つづりの揺れなどに容易に対応することができ、様々なビジネスチャンスの可能性を提供する。

本研究成果の副産物として、実証研究で対象にした 24 以上の言語の機械辞書がある。これらの機械辞書は、本研究費により可能となった民間出版社からの電子辞書利用許諾および本研究に従事した研究者の同僚・友人たちの好意によるところが大きい。ここに記して謝辞を述べたい。

【誌上発表リスト】

[1] Kayo IKEDA, Hideho NUMATA, Masakatsu KANEKO, Kazuhiko MACHIDA, “Proposal for a Multilanguage Text Input Support System that is Easy for Beginner Language Learners”, Proceedings of the 3rd International Universal Communication Symposium, pp.109-114., (December, 4, 2009)

[2] Kazuhiko MACHIDA, “Indian language resources and Unicode encoding - Some Problems of Hindi texts in Devanagari -“, 31st ALL INDIA CONFERENCE OF LINGUISTS, pp.92-94., (December, 16, 2009)

[3] 町田和彦、「インドの英語」『英語世界のことばと文化』成文堂、pp.162-175 (2008)

【本研究開発課題を掲載したホームページ】

http://www.aa.tufs.ac.jp/~kmach/scope/scope.htm