E-7

PostgreSQL と JSP を用いた 多言語データベース検索アプリケーションの構築

The Application Construction for Multilingual Database Using PostgreSQL and JSP System

堀　一成† 前田　彩＊ 石島　悌‡
Kazunari Hori Aya Maeda Dai Ishijima

1．まえがき

コンピュータによる多言語間の自動翻訳や自動通訳を実現するためには、その基盤となる多言語間の言語データベースの整備が不可欠である。すでに大阪外国語大学では、常用基本単語５０００語のリストとして中国語・タイ語、また約１０００文で構成される旅行会話データ集として中国語・ビルマ語・タイ語・ベトナム語・ヒンディー語をテキストデータとして作成している。さらに平成１４年度中に数言語のデータ追加を行う予定である。加えて各言語のテキストデータに対応した音声データについても蓄積作業を開始している。このような多言語マルチメディアデータを速やかにデータベース化し、広く活用してもらえるようWebアプリケーション等の整備を行わなければならない。

フリーのデータベースである PostgreSQL[1]は、多言語情報を手軽に扱える UTF-8 文字コード[2]の処理をサポートしている。また、音声や映像データなどの大きなバイナリデータも、ラージオブジェクトという形で扱うことができる。これらの特長に着目し、これまで我々は簡易な多言語データベースを作成し、さらに検索 Web アプリケーションを作成したことを発表してきた。この Web アプリケーションは Perl ベースの CGI[3]または、PHP ベース[4]のものであった。

前記のさまざまな種類の言語データを活用し、将来、システムを発展させて自動翻訳や自動通訳といった大規模なシステムに拡張することが、本研究の最終到達目標である。そして、そのような大規模システムを開発するためにはオブジェクト指向的開発手法が適用できるサーバサイド Java システムを用いることが望ましいと考えられる。そこで今回、その開発の第一歩として JSP と JDBC ドライバを用いたシステムを構築した。

2．システムの概要

図１にシステムの概要を示す。ユーザが UTF-8 対応 Web ブラウザから多言語単語データベースのページにアクセスし、入出力言語と入力言語の単語リストから必要な単語を

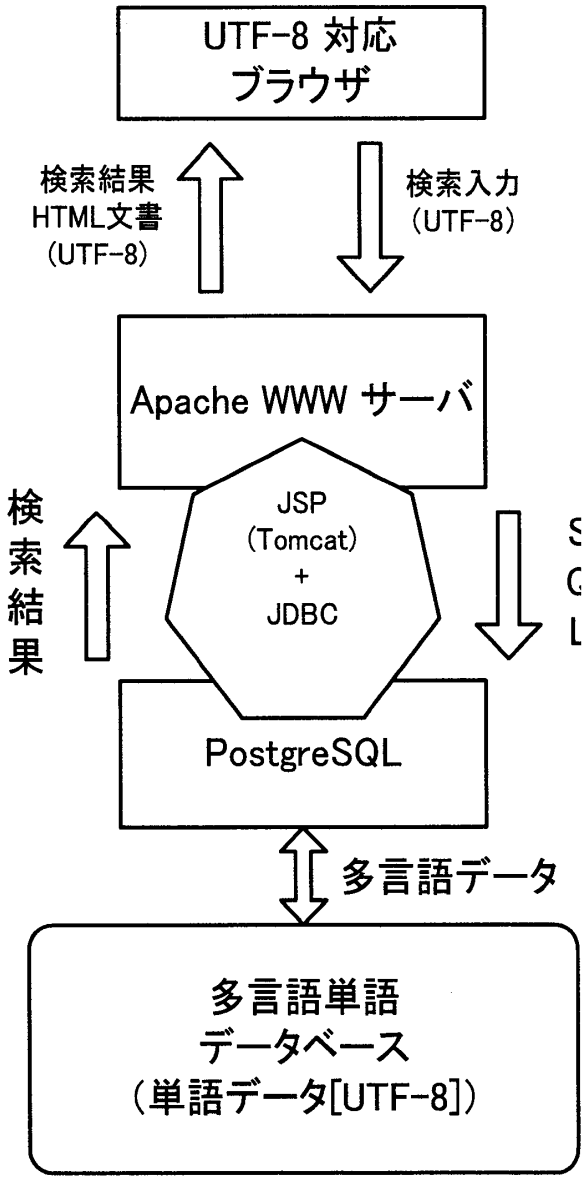

図１　システムの概念図

†大阪外国語大学, Osaka University of Foreign Studies
＊ TIS 株式会社， TIS Corporation
‡大阪府立産業技術総合研究所,
Technology Research Institute of Osaka Prefecture

選択すると、JSP 中に埋め込まれた JDBC ドライバが PostgreSQL データベースにアクセスし、検索結果を得る。以上の行程において扱われる文字データ、スクリプトの記述文字コード全てに UTF-8 文字コードを採用している。特に Java 言語は UTF-8 文字コードと親和性が高く、ほとんどの場合スクリプト記述に当たって扱うデータの多言語性に特に注意を払う必要はなかった。ただし、HTML の SELECT メニューにより入力した文字列データについては UTF-8 データを正常に扱えず、

```
String inputWord = request.getParameter("input_word");
inputWord = new String(inputWord.getBytes("8859_1"), "utf-8");
```

と文字化け対策のコードを記述する必要があった。

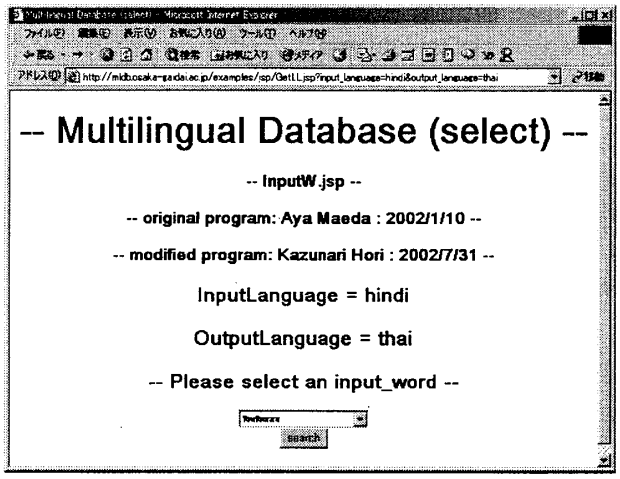

図 2　Web アプリケーション（単語選択時）のスクリーンショット

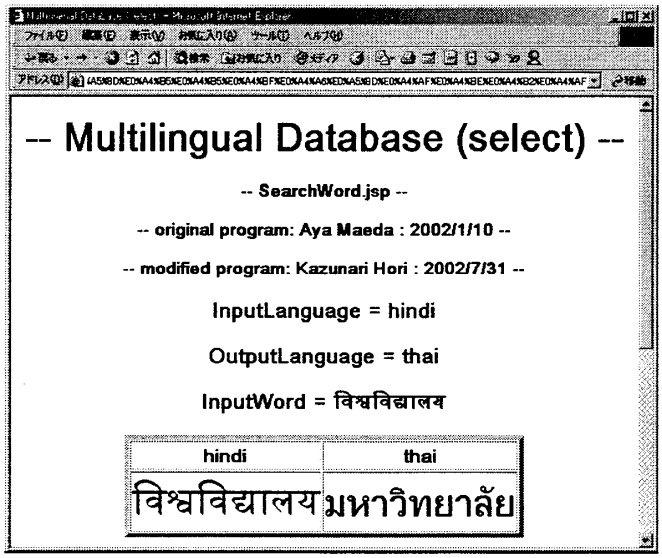

図 3　Web アプリケーション（検索結果表示）のスクリーンショット

図 2 は検索入力言語と出力希望言語を指定した後、検索する単語を input_word として入力している画面である。この図の場合、検索入力言語はヒンディー語、出力希望言語はタイ語を選んでいる。図 3 は検索結果を表示した画面である。ヒンディー語、タイ語ともに日本語で「大学」を意味する単語を表示している。

3. 今後の課題

（1）現在、より柔軟なシステム開発が行えるよう、MVC モデルアーキテクチャに沿ったシステムに移行作業を進めている。進展があれば発表時に併せて紹介する予定である。

（2）ラージオブジェクト機能を用いた、音声データ提供アプリケーションを作成する。併せて i-mode 携帯電話等ポータブルシステムに対応するアプリケーションを作成する。（いずれも PHP ベースでは既に開発済み[4]）

（3）複雑なタグ付け情報を必要とする言語解析システム用データは、単純な表形式の関係データベースでは十分にあらわすことができないと考えられる。単語、例文データのタグ付 XML データ化とそのデータベース化が必要である。望ましいデータ形態の研究と、XML データベースシステムおよびアプリケーションの開発を行う。

謝辞

本研究は、科学研究費補助金　基盤研究（B）『多言語同時処理によるアジア系言語の自然言語翻訳に関する基礎研究』課題番号 14310220 に基づく研究である。

参考文献

[1] 石井達夫: PostgreSQL 完全攻略ガイド 改訂第 3 版, 技術評論社 (2001)

[2] 錦見美貴子、高橋直人、戸村哲、半田剣一、桑理聖二、向川信一、吉田智子：マルチリンガル環境の実現，プレンティスホール，(1996)

[3] 堀一成、石島悌: "PostgreSQL による多言語単語データベースの構築"　情報処理学会　第 62 回全国大会

[4] 堀一成、石島悌:"PostgreSQL を用いた多言語文字・音声データベースの構築とアプリケーションの開発"
情報処理学会　第 63 回全国大会