文字コード独立の多言語テキストeditorの実装

5R－10

張暘　梅村恭司

豊橋技術科学大学 情報工学系

1．はじめに

文字コードと符号化方式が多種多様にあり、これらを統一的に処理するためには、OSを国際化するべきであり、Unicodeの採用はその手段のひとつであると考えられる。しかしながら、それは様々な要因で、まだ広い範囲でサポートされていない。また、たとえUnicodeを使っても、異なる文字コードを同時に表示させることができないという問題がある。

本研究では、文字コードに依存したテキスト文字を文字コードに依存しないGIFファイルに対応づけることで、文字コードに独立したHTMLファイルに変換するツールを作成した。このHTMLファイルを元のテキストファイルと同じインターフェースで表示しながら編集できるeditorを作成することで、事実上の多言語テキストeditorを実現した。また、文字コードセットを拡充でき、OSではサポートされていない言語にも対応することができる。

2．システムの概要

システムの開発では、Java言語を用いる。LinuxとWindows95およびWindows98では、システムが正しく動作することが確認できた。システムが対応している文字コードセットは日本語（JIS、SJIS、EUC）と中国語（GB、HZ「簡体字」とBIG5「繁体字」）と英語（ASCII）である。このシステムはBdxToGifクラスとTxtToHtmlクラスの二つツールおよび多言語テキストeditorから構成される。BdxToGifはLinux上の各コード体系のBDFファイル（テキストファイル）から各コード体系のGIF libraryを作るツールである。TxtToHtmlは各コード体系のテキストファイルをHTMLファイルに変換するツールである。

3．実現方法

多言語テキストeditorを使うには、まず、BdxToGifを使って、各コード体系のGIF libraryを作っておく必要がある。新たなニーズによって、新しいコード体系が必要になれば、このツールを使って、GIF libraryを追加すれば良い。TxtToHtmlを使うと、変換対象になる各テキストファイルのコード体系に対応する変換クラスが呼び出されて、HTMLファイルが作成される。これは、文字コード体系に依存するテキストファイルの文字を文字コード体系に依存しないGIF形式の画像ファイルに対応させたものである。多言語テキストeditorをマルチウィンドウで実現し、いくつかの必要なHTMLファイルを多言語テキストeditorの各ウィンドウにロードして、多言語テキストeditorの編集機能を利用し

Implementation of multiple languages' text editor independent of character encoding

Yang Zhang and Kyoji Umemura

Department of Information and Computer Sciences, Toyohashi University of Technology

て、元のコード体系に依存しない多言語 HTML ファイルが作成できる。(多言語テキスト editor で表示されるのは、元のテキストと同じだが、実際に編集されるのは HTML ファイルである。) また、多言語テキスト editor は HTML ファイル以外に X-bitmap 形式の画像ファイルでも出力することができる。

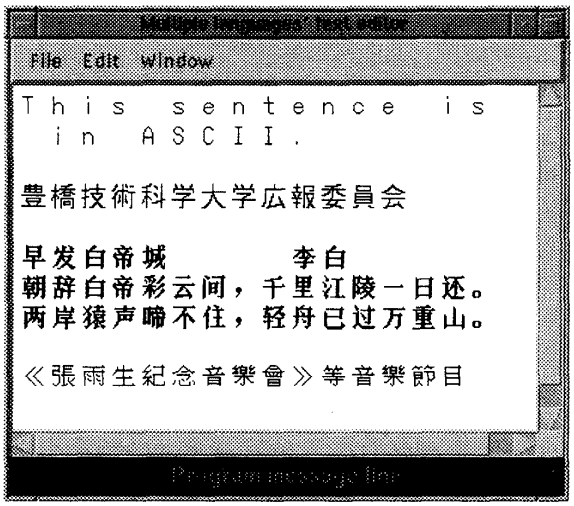

図1

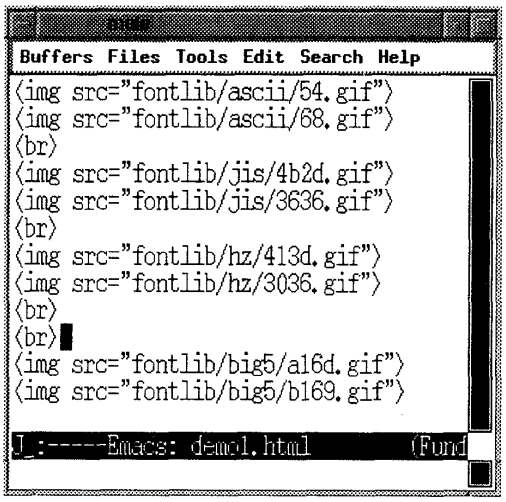

図2

実行例として、多言語テキスト editor の編集機能を利用して作られた多言語 HTML ファイルを表示した例を図1に示す。この HTML ファイルは英語、日本語、中国語 GB（大陸、シンガポールなどで使われた「簡体字」)、中国語 BIG5（台湾、香港などで使われた「繁体字」）の文字を含んでいる。図2はこの HTML ファイルのコードの一部分である。

4．システムの性能

本システムは GIF library を使うので、ハードディスクに多くのスペースが要求される。(ひとつ GIF ファイルのサイズは１００バイト以内である。) 作られた HTML ファイルのサイズも大きいである。表示するときに、ハッシュテーブルを使って、必要な GIF ファイルだけをメモリに読み込むので (ハッシュテーブルに登録した GIF ファイルをロードしない)、表示の実行時間はテキスト中の文字の種類の数に依存する。また、システムの性能を向上させるには、多量のメモリが必要である。実験では、Java インタプリターのヒープメモリを多く指定すれば、実行時間は半分以下になった。市場の主流マシンでは、十分実用に耐えるものである。

5．おわりに

本研究では、我々は文字コード独立の多言語テキスト editor を実装した。これにより、多言語テキストを統一的に表現することが可能になった。現在、本システムは日本語（JIS、SJIS、EUC）と中国語（GB、BIG5、HZ）と英語（ASCII）に対応しており、対応するコードがない文字も処理できる。

参考文献

「1」Ken Lunde：日本語情報処理，ソフトバンク株式会社（１９９５）
「2」有賀 妙子・竹岡 尚三： Java1.1 プログラミング，ソフトバンク株式会社（１９９７）