3R-7

ローマ字転写による伝統的モンゴル語入力方式の構築

バトモンク　　中平　勝子　　三上　喜貴

長岡技術科学大学

1. はじめに

モンゴル語の表記には伝統的モンゴル文字(以下単にモンゴル文字と表記)とキリルモンゴル文字(以下キリル文字と表記)が使われており、モンゴル国ではキリル文字を、中国内モンゴル自治区では伝統的モンゴル文字を主として用いる。モンゴル文字はもともとセム系言語が用いていた文字体系（アラム文字）を出発点として形成された文字であり、全く音韻体系の異なる言語から「借用」した文字体系であるために、表音文字でありながら音と形が一対一対応でなく、異音同形問題が存在する[1]。また、アラビア文字と同様、同一文字であっても語中の位置によって表示形が異なる。こうした事情のため、文字コード化に際しても困難が多く、国際標準[2]として制定されたモンゴル文字コードはかなり複雑な構造をしている[3]。

また、国際標準文字コードが成立しているにもかかわらず、これを用いた標準的な入力システムが普及しておらず、多数の特殊な入力方式・特殊な文字コードが依然として用いられている現状にある[4]。

そこで、筆者らは、簡単で、どんな環境でも使えるようなモンゴル文字の入力方式の 1 手法として、ローマ字転写による入力方式を構築した。

2. 入力方式の提案

モンゴル文字の電子情報の互換性を前提とした国際標準に基づく入力システムが単純入力方式であり、ユーザの負担が大きいため、効率の改善が必要とされている。ここで単純入力方式は単語を入力する場合、物理的に一文字ずつ入力する方式と定義する。

どのような環境でも使えるためには標準キーボードに基づく入力システムが必要であり、一方で、多くの一般の利用者にとって簡単であるためには、日本の仮名漢字変換システムのようにローマ字入力を可能とした上で、後方文字を予測し、候補文字列を選択する予測入力機能が効果的である。

本稿では、ローマ字転写によるモンゴル語予測変換入力方式を提案する。本方式の有用性を確認するため、以下の二つのステップを踏まえて、分析を行った。

2.1 単語頻度分布

モンゴル語入力方式の構築にあたり、モンゴル語の計量的特徴をつかむために、モンゴル語 4000 単語のコーパス[5]を用いて、単語頻度分布を測定した。べき乗則を用いて、単語頻度分布の分析を行った結果、モンゴル語にとって予測変換入力が可能であると推測した[6]。

Development of Romanized Input Method for Traditional Mongolian Script

Batumengke, , Katsuko T Nakahira , Yoshiki Mikami
Nagaoka University of Technology

2.2 入力コスト評価

次に、ユーザの入力負担を軽減できるかどうかを検討するため、入力コストの比較を行った。その評価方法は単純入力方式と予測変換入力方式を使用したときの、平均ストローク数を対象にする。

本稿では以下の二つの式により、先述のコーパスに含まれる同一先頭文字で始まる単語リストを用いて、計算を行った。式(1)は単純入力システムの入力コストを計算した。

$$\sum_{i=1}^{n} PiLi = C \qquad \cdots (1)$$

P_i：i 番目の単語の出現確率
L_i：i 番目の単語の長さ
n ：単語数
C ：入力コスト

予測変換入力システムの入力コストを式(2)で計算した。

$$\sum_{i=1}^{m} Qi \sum_{j=1}^{ni} PijMj + I = E \qquad \cdots (2)$$

I ： 先頭文字の入力コスト
Q_i： i 番目の文字で始まる単語出現確率
P_{ij} ： i 番目の文字ではじまる単語集の j 番目の単語の出現確率
M_j ： 順位
m ： 先頭文字の種類
n_i ： i 番目の文字ではじまる単語の総数

上記の二つの方式を用いて計算した結果を表 1 に示す。

表 1 には先頭文字を固定した単語数とそれの入力コストを示す。入力コストが C > E の場合、予測変換入力方式が有利である。表ではほぼ 8 割の入力コストが C > E という結果になっている。これより、予測変換入力のほうが有利であると考えられる。

表 1　入力コストの比較表（抜粋）

先頭文字	含有単語数	C	E	先頭文字	含有単語数	C	E
ᠨᠠ	609	6.51	11.38	ᠠ	246	5.92	2.64
ᠣᠨ	388	6.46	7.14	ᠣᠯ	265	6.59	2.73
ᠣᠷ	406	5.74	5.67	ᠣᠴ	186	3.96	2.12
ᠣᠨ	339	6.54	3.87	ᠣᠨᠠ	182	5.23	2.13
ᠣᠯ	279	3.71	3.26	ᠣᠷ	196	3.62	2.15

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.

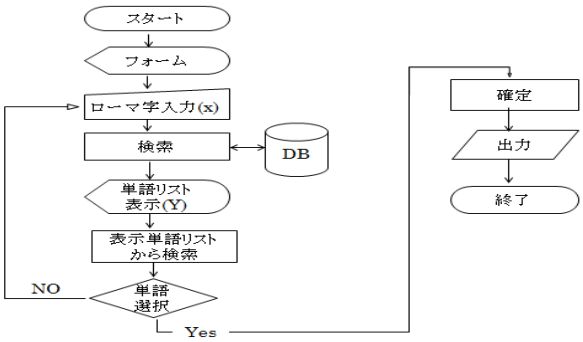

図 1 システムの全体図

3. 入力方式の実装

本稿では Ajax[7]を取り入れたＷｅｂアプリを作成した。MySQL を使って、サーバ側に辞書 DB を作製し、ユーザが入力した文字をサーバに送り、サーバから検索した結果を返す。今回は Ajax を使うことで、非同期的にサーバと通信することで、検索結果を常時に表示し、選択することを可能とした。また、単語頻度データを持つ辞書を引用し、各文字で始まる単語リストをまとめ、それに頻度情報を加え、ユーザが一文字を入力したとき、その文字で始まる使用頻度の高い単語を優先的にリストアップすることで、入力コストの向上を図った。図 1 はシステムの全体流れを示す。

DB の構造として、コーパスをローマ字に変換したデータを基本とし、ローマ字転写、モンゴル語、出現頻度を記録した。しかし、すべての単語を辞書に載せることが不可能なため、DB に存在しない単語を入力する際には単語内の全文字を入力することで入力を行い、DB に登録する。

モンゴル文字の入力を行うには、次の問題がある。(1)表示形の選択を行うとき、特殊記号(MVS,$FVS_{[1-3]}$)を使用している。また、(2)分かち書き助詞をそれと同一文字で構成される他単語と区別するため、NNBSP(NARROW NO-BREAK SPACE)を使用する。どちらも等確率で出現するため、ユーザが挿入する特殊記号の使用頻度が多いとき、効率的な入力に影響を与える。それで、(2)は助詞と一体として、辞書登録によって解決できるので、本システムで(2)を実装した。(1)はユーザが挿入する形になっているが、自動で行うにはルール化が必要である。それで本稿では(1)を検討してない。

4. 評価実験

本入力システムを実装し、具体的に入力効率を向上することが可能であるかどうかを検討した。本入力方式を使って実験をおこない、それと同時にアンケート調査も実施した。実験方法として、実験者 10 人を 2 グループに分けて、マイクロソフト OS に付属する IME(A)と本アプリ(B)の二つの入力方式でそれぞれ限られた時間内に 2140 文字からなる文章を入力してもらった。二グループに分けて、入力順序を変えることによって、バランスの取れたデータを収集できると考えた。システムＡは国際標準に対応した入力システムであり、Vista 以降の Windows では、UCS/Unicode 用のモンゴル文字フォント（Mongolian Baiti）と IME が実装されているため、今回の実験に採用した。

表 2　実験結果

	入力システム	総入力文字数		
		総被験者	IME 非利用者	IME 常時利用者
平均（全体）	A	562	473	919
	B	551	529	640

表 2 に示すとおり、総被験者の結果により、システムＡとＢの入力文字数がほぼ同じである。IME 非利用者を対象にした場合、システムＢの方は 10%程入力文字数が多いことがわかる。IME 利用者にとってシステム A のほうが有利である。

これにより、予測変換入力方式を使うことで入力速度を向上させることが可能であると考えられる。

5. まとめ

本稿ではモンゴル語の単語頻度情報を持つコーパスを用いて、入力コストを比較し、入力高速化する可能性を示した。また、ユーザの負担を軽減できると考え、キーボード依存の少ないローマ字転写による縦書きモンゴル語入力方式を構築した。このアプリを用いて実験を実施し、その結果から予測変換入力方式が効果であるということを確認できた。

今後の課題として、モンゴル語の特徴を取り入れた入力を目指す。また、ユーザビリテイの向上についても検討の必要があると考えている。

参考文献

[1] 确精扎布 ：<< 蒙古文编码 >>, 内蒙古大学出版社, 2000 年
[2] [Unicode Standard 6.0, p.426]。
[3] Traditional Mongolian Script in the ISO/IEC 10646 and Unicode Standards. Myatav Erdenechimeg, Richard Moore and Yumbayar Namsrai. UNU/IIST Report No. 170
[4] A Study of Traditional Mongolian Script Encodings and Rendering: Use of Unicode in Open Type fonts. International Journal on Asian Language Processing 21(1):23-43. Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, Fuminori Kimura, Akira Maeda
[5] Mongolian Speech Corpus, http://www.panl10n.net/english/OutputsMongolia2.htm#Linguistic Resources 、accessed Jan. 14th 2011.
[6] ローマ字転写による縦書きモンゴル語入力方式構築のためのモンゴル語文字頻度分析. 巴図孟克, 中平勝子, 三上喜貴. 情報処理学会第 73 回全国大会
[7] 基礎 Ajax + JavaScript. 佐藤和人

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.