1S-7

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

バトモンク　　中平　勝子　　三上　善貴
長岡技術科学大学

1. はじめに

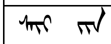
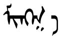
高度情報化時代を迎え、モンゴルにおいてもメール等による言語テキストの電子的交換が日常的に行われるに至っている。その際、平易な言語入力方式を構築することが利用者の負担を減らす上で必要の課題であるが、モンゴル文字の持つ特殊な構造により平易な入力が実現しておらず、特殊な入力方式・特殊な文字コードに頼ることが多かった。

経済成長、電子化の発達に伴い、若い世代を中心にいろいろな情報をインターネットから収集する動きが活発になっている。モンゴル文字の電子化が英語と日本語に比べ、やや遅れている状況にある。

それは、モンゴル文字の特徴に起因している。モンゴル語は伝統的モンゴル文字(以下モンゴル文字と表記)とキリル文字の現代モンゴル語(以下キリル文字と表記)が使われ、モンゴル国ではキリル文字を主に、中国内モンゴル自治区で伝統的モンゴル文字を主に用いている。モンゴル文字は母音文字が 8 文字、子音文字が 27 文字で合計 35 文字からなる[1]。モンゴル国では 1946 年からキリル文字が公式に使用される[2]。キリル文字に、母音文字が 13 文字、子音文字が 20 文字、記号 2 文字の 35 文字からなる。伝統的モンゴル文字はキリル文字に比べて汎用性が低く、構造をつかみにくい(表 1 を参照)。そのため、電子化には向きにくかった。最近ハルタルフーら[3]伝統的モンゴル文字のディジタル図書館の構築などモンゴル文字の電子化活動が活発に行われて、それが学校教育、情報機関などに日々求められている。

そこで我々は使い易いモンゴル語入力方式を実現するため、予測入力支援によるモンゴル語文字頻度分析を行った。

表 1. 伝統的モンゴル文字とキリル文字の例

伝統的モンゴル文字	キリル文字
	аюул
	саяын
	Мянга

2. 予備分析

モンゴル文字入力方式の構築に当たり、モンゴル文字の計量的特徴をつかむために、A 氏が蒙古語新聞から収集した約 80-90 万の単語頻度表データ[4]のうち公開されている上位 100 語を用いてどの程度の負担軽減や入力高速化が期待できるかを分析した。

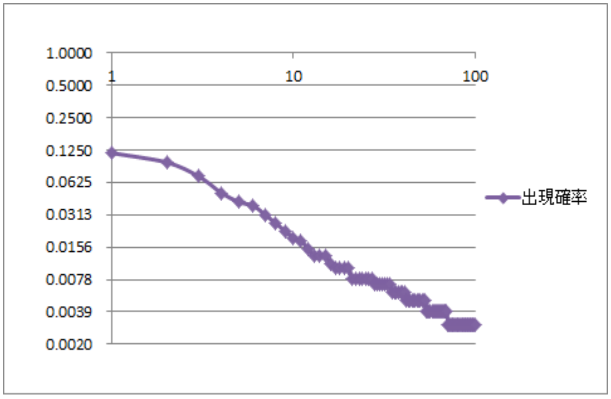


図 1. 単語の出現頻度データ

図 1 には横軸に単語頻度の順位 K と縦軸にその単語の出現確率を両対数で示す。このグラフは $K^{-\alpha}$ に比例する。自然言語処理においては、順位 K と頻度を両軸とする両対数グラフは Zipf 法則に従うとされ, 図 1 では $\alpha \sim 0.86$ であり、Zipf 法則($\alpha=1$)に概ね合致する。

自然言語処理の方法でモンゴル語の特性を調べることが可能であると考えられる。

3. 頻度分析

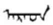
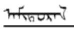
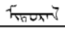
表 2 に同一文字で始まる単語頻度分析を示す。

「Frequency Analysis For Construction Mongolian Input Method By Employing Latin Characters」

†「Batumengke
Katsuko T Nakahira
Yoshiki Mikami ・
Nagaoka University of Technology」

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

表 2.　同一文字で始まる単語

単語	先頭文字列	後続文字列	出現確率(P)	単語長(N)	C=P*N	出現順位(重み)(M)	E=P*M
			0.0100	4	0.0400	1	0.0100
			0.0080	3	0.0240	2	0.0160
			0.0060	4	0.0240	3	0.0180
			0.0050	4	0.0200	4	0.0200
			0.0040	6	0.0240	5	0.0200
			0.0030	8	0.0240	6	0.0180
					0.1560		0.1020

先頭文字が同じ単語を分類し、出現確率降順に並べ、それと同時に単語長を求めで、単語入力に使う相対的な文字入力量の合計を出す。一方、単語の確率の高い順から並べ、その出現順位を決め、そこから先頭文字が同一単語の相対的な文字入力量の合計を出す。この二つを比較することにより、労力の軽減を簡単に求めることができる。表 2 から同一文字で始まる単語の出現確率の高い順に並べることにより、入力速度を向上させることができると考えられる。

4. 予測変換を取り入れた入力の負担軽減

本稿では、モンゴル文字入力の負担軽減効果をはかるために、コーパスを先頭文字[注 1]ごとに分類し、それぞれの出現確率と順位の関係を調べる。1 文字目の情報に基づいて入力メニューを提案する場合、第 1 文字で出現確率が同じである場合、第 2 文字の出現確率の高い順位で決めることにしている。

ここでは出現確率(Pw_i)は順位 K の$-\alpha$乗に比例すると仮定する。$\alpha=1$ ならば、Zipf 法則に従い、バランスが取れたコーパスを示す。$\alpha>1$ の場合、文字の出現確率の差が大きくなり、次の文字の特定が易くなる。それと比べ、$\alpha<1$ の場合、文字同士の出現確率が小さくなり、この方式に従い、後続文字の特定が難しくなると推定される。先頭文字ごとに分類した K- Pw_i の関係をべき乗近似で求めた際のαを表 3 に示す。

表 3 の左半分は$\alpha>1$ で、後続単語を簡単に特定できると考えられる。αの値が 1 に近い範囲であれば、平均的な文字列になっている。表 3 の右半分は$\alpha<1$ で後続文字の種類が多く、ほぼ等確率で出てくるため、先頭文字にした文字列のα値が比較的に低く、後続文字の特定がむずかしくなると推定される。このように分析した結果からモンゴル文字列とモンゴル語コーパスの特徴を表すことができる。

さらに、綿密なモンゴル文字頻度分析によって、モンゴル語入力の負担軽減を評価できる。

表 3.　先頭文字列ごとのα値

先頭文字	α	先頭文字	α
	1.520		0.646
	1.358		0.590
	1.109		0.579
	0.949		0.525
	0.870		0.440
	0.807		0.402
	0.800		0.345
			0.248

5. まとめ

本稿では、モンゴル語の単語頻度を Zipf 法則に基づいて、分析した結果、出現確率の高い単語を先頭に並べることで、入力高速化する可能性を示した。また、出現確率と順位の比例により、ユーザーの負担を軽減できると考えられる。

今後の課題として、適切なコーパス[6]を用いてモンゴルの文字の特性を表現することである。最終的に、モンゴル文字列の並びや単語の遷移確率をまとめ、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することである。

参考文献

[1]涛高, モンゴル語文法概要、内蒙古少年児童出版社、通遼市, 1986. (モンゴル語)

[2]満 都拉、藤井 敦、石川 徹也、伝統的モンゴル語の電子化方式とテキスト検索への応用、電子情報通信学会論文誌、Vol. J88-D-II　No. 10 pp. 2102-2111, 2005

[3]ハルタルフー・ガルマーバザル、伝統的モンゴル文字文書のディジタル図書館に関する研究、アート・リサーチ 8, 33-43, 2008-03 立命館大学

[4]http://menkdev.blog.163.com/blog/static/116658303200951152712204/

[5]Steven Bird, Ewan Klein, Edward Loper, 入門自然言語処理

[6]Mongolian Speech Corpus, http://www.panl10n.net/english/OutputsMongolia2.htm#Linguistic_Resources_、accessed Jan. 14th 2011.

「[注 1]:モンゴル文字「ᠩ(ng)」が単語の先頭に入らない」

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.