伝統的モンゴル語と現代モンゴル語の双方向的な翻字手法

満　都拉　　藤井　敦　　石川　徹也
筑波大学大学院図書館情報メディア研究科
{mandula, fujii, ishikawa}@slis.tsukuba.ac.jp

１．はじめに

現在使われているモンゴル語には, 縦書きモンゴル文字を用いて表記する「伝統的モンゴル語」とキリル文字を用いて表記する「現代モンゴル語」がある．伝統的モンゴル語は主に中国の内モンゴル自治区で使われており, 現代モンゴル語は主にモンゴル国で使われている．

政治的な隔離に伴い，モンゴル国では 1941 年に縦書きモンゴル文字を廃止したため, モンゴル国で伝統的モンゴル語を読み書きができる人は少なくなっている．内モンゴル自治区では，現代モンゴル語は普及しなかったため, 現代モンゴル語を読み書きできる人が少ない．そこで，モンゴル国と内モンゴルの間で情報交換が困難になっている．

これら 2 つのモンゴル語は, 使用する文字の体系が異なるだけで, ほぼ同じ文法規則に基づいており，どちらも表音文字を使用する．

本研究は，これらの性質に着目し，一方のモンゴル語で書かれたテキストを他方のモンゴル語に文字単位で変換するための翻字手法を提案する．具体的には，母音や子音といった音素に関する対応規則と正字に関する規則を作成して翻字を実現する．

その結果, モンゴル国と中国の内モンゴル間での双方向的な情報交換に貢献することができる．

伝統的モンゴル語は文字構造の特殊性が原因でテキストの電子化が進んでおらず, テキスト処理に関する研究が遅れている．それに対して，現代モンゴル語の電子化テキストは普及している．そこで, 現代モンゴル語の電子化テキストを伝統的モンゴル語に変換して, 伝統的モンゴル語のテキスト処理研究を促進する効果もある．

以下，２．で先行研究について検討し，３．で本研究の翻字手法について説明する．４．で本翻字手法の評価を行う．５．で翻字結果を利用したモンゴル語の言語横断検索システムを提案し，６．で検索システムの評価実験について説明する．

２．先行研究

中里ら[2]は，伝統的モンゴル語の単語を現代モンゴル語に変換する手法を提案した．この手法では，１つモンゴル語の単語を伝統的モンゴル語と現代モンゴル語にそれぞれ変換単位として分割するとき, 切り分ける個数が等しいかどうかで判断している．

しかし，伝統的モンゴル語では伝統を守り，習慣上の表記によって音節上必ずしも一致しない場合がある．例えば，日本語では「遺言」を伝統的モンゴル語で「geriyesu」のように 4 音節で表記する．しかし，現代モンゴル語では「gerees」のように 2 音節で表記している．

そこで, 同一単語を音節により両方を等しい単位で切り分けることは困難である. 切り分け単位が等しくなれば正しくとする判断基準はモンゴル語の両表記法の変換単位として相応しいかどうかは検討する必要がある．

中里らは現代モンゴル語から伝統的モンゴル語への変換はしていない．また，変換対象は単語単位であり，文章（テキスト）ではない．

３．本研究で提案するモンゴル語の翻字手法

3.1　概要

本研究の翻字手法は，音素単位で変換し，変換先言語の正字法を適応して表記を特定する. 正字法で特定できない場合は例外として処理する. さらに, 統計的言語モデルを用いて変換誤りを自動修正する．

3.2　現代モンゴル語から伝統的モンゴル語への翻字

現代モンゴル語から伝統的モンゴル語への翻字手法の概要を図１に示す．

伝統的モンゴル語と現代モンゴル語の大きな違いは，助詞の分かち書きと母音の弱化である．

伝統的モンゴル語では，助詞を分かち書きし，分かち書きされた助詞は単語の性（陽性か陰性）によらず決められた字形を持つ. 例えば，日本語の「私達の」は「  」と表記し，「彼らの」は「  」と表記する．

しかし, 現代モンゴル語は助詞を分かち書きしない場合が多く, 単語の性によって母音調和規則

に従い表記が変わる．例えば,「私達の」は陽性語なので「манай」と表記し，助詞「の」は陽性の「ай」と表記する．しかし,「彼らの」は陰性語なので「тэдний」と表記し，助詞「の」は陰性の「ий」と表記する．このように単語の性により助詞の表記が変わる．

そこで, 現代モンゴル語から伝統的モンゴル語に変換する時, 助詞を正しく分割し，単語の性を特定することが重要である．そして，特定した単語の性と伝統的モンゴル語の正字法に基づいて弱化母音を補正し，字形を決める．

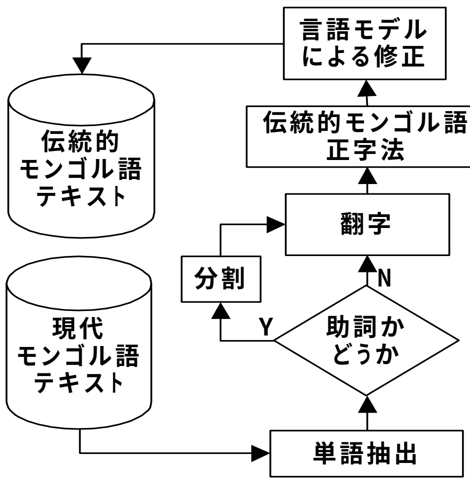


図１　現代モンゴル語から伝統的モンゴル語への翻字

本手法では，аас、аар、ийг のような助詞を 30 個用意し, 単語の語尾をこれらの助詞と照合して分割する．

母音弱化に対しては, 伝統的モンゴル語の正字法[1,3,4]に従い, 変換した結果における弱化母音の補正を行う．

しかし, 伝統的モンゴル語は 800 年の歴史を持ち，正字法が語形論を重視し，伝統を守る規則があるため, 現代モンゴル語の発音と合わない場合がある．例えば,「人間」を伝統的モンゴル語では「humun」と表記する．しかし，現代モンゴル語では「хҮн」と表記する．このような「mu」が「Ү」になる表記上不規則の特殊な語を対応辞書を作成して対応する．

また, 伝統的モンゴル語では同音字を区別するために弱化母音を字形で区別する．例えば，発音[hair]を伝統的モンゴル語で「愛」の意味を表すときは「ᠬᠠᠢᠷᠠ」（haira）と表記し,「砂利」の意味を表すときは「ᠬᠠᠢᠷ」（hair）と表記する．

しかし，現代モンゴル語では，音声学論を重視し，弱化母音を表記しないため，現代モンゴル語では両方を「хайр」表記し「愛」なのか「砂利」なのか区別できない. このような語の翻字は文脈情報が必要なので，本手法では区別せず，今後の研究課題とする．

3.3　伝統的モンゴル語から現代モンゴル語への翻字

3.2 で説明したように，伝統的モンゴル語と現代モンゴル語の違いから, 分かち書きされる伝統的モンゴル語の助詞を判断し, 助詞を単語に接続する語尾処理が必要になる. 翻字手法の概要を図２に示す．

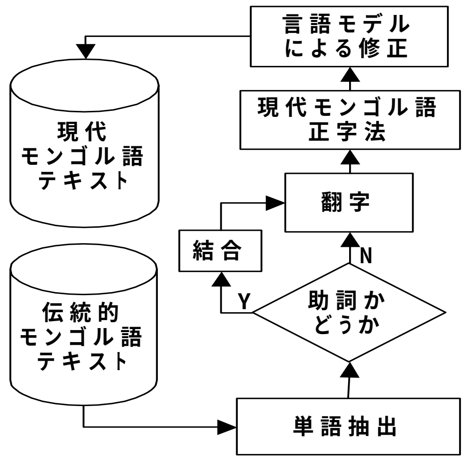


図２　伝統モンゴル語から現代モンゴル語への翻字

助詞かどうかの判断は, 文頭から連続している２つの単語を抽出し，二番目の単語が助詞かどうか検査する．二番目が助詞の場合，一番目の単語と連結して，次の１つの単語を調べる．二番目が助詞でない場合, 一番目の単語を１つの単語として確定する．次に 2 番目の単語とその次の単語を取り出し, 二番目の次の単語が助詞かどうかを調べる．

次に，母音に挟まれたｇ，ｙ，ｂ，ｍが長母音であるかどうかを調べる．例えば,「agasi、ariyatan、debel」の場合，ｇ，ｙ，ｂで構成された音節が長母音になる．しかし,「hagas、ebugen」のｇ，ｂ で構成された音節は長母音にならない. このことを自動的に判断することは困難なので，人手で辞書を作成して対応する．

次に，文字単位で変換し，キリル文字表記の正字法[4,6]を適応することによって単語を構成する．

さらに, 現代モンゴル語コーパスから作成した統計的な単語ユニグラムモデルを利用して, 変換精度を向上させる．まず，ユニグラムに含まれな

い単語を誤変換として検出する．次に，ユニグラム中の単語との編集距離を計算し，編集距離が最も短い単語に修正する．ただし，同順位の単語が複数ある場合は，ユニグラム頻度が大きい単語を選択する．

現代モンゴル語では，固有名詞や外来語の語頭の一文字を大文字で表記する．しかし，伝統的モンゴル語では大文字小文字の区別がないため，本手法は固有名詞を判断することができない．

４．翻字手法の評価

現代モンゴル語から伝統的モンゴル語への翻字はWebサイトhttp://onigoo.olloo.mnから入手した新聞記事10件（述べ6,943語，異なり2,223語）を実験データとして利用した．

変換結果は，延べ語数で1,253語が間違っており，正しく変換されたのは5,690語だった．この結果，正解率は82.0%である。

異なり語数で356語が間違っており，1,867語が正しく変換された．この結果，異なり語の正解率は84.0%だった．

しかし，変換誤りがあっても内容理解を防げることはなかった．

伝統的モンゴル語から現代モンゴル語への翻字は，人手で入力した新聞記事５件（述べ1,278語，異なり730語）を利用して評価した．

変換結果は，述べ語数 427 語が正しく変換された．正解率は 33.4%（427/1278）である．異なり語数で 243 語が正しく変換された．正解率は33.3%（243/730）である．

統計的言語モデルを用いて変換誤りを自動検出した．その結果，再現率は50.6%（123/243）で，精度は93.9%（123/131）だった．検出に失敗した誤変換は 8 語あった．この 8 語を分析した結果，略語の大文字表記の文字列が正字法違反の文字列に表層一致した結果であった．

ユニグラムに含まれなかった 117 語の誤りを自動修正する実験も行った．その結果，自動修正の再現率は59.5%（50/84）であり，精度は63.3%（50/79）であった．

５．モンゴル語の言語横断検索システム

３．の翻字手法で構築した伝統的モンゴル語のテキストデータを応用し，モンゴル語の言語横断検索システムを実現した．システム構成を図３に示す．

図３において，入力インタフェースと出力インタフェースは筆者らの提案した入出力インタフェースである[5]．ユーザが端末から検索質問をローマ字で入力する．この入力された検索質問が出力インタフェースによってモンゴル語で表示される．同時にローマ字形式の入力情報が検索処理に渡される．

検索処理は，ユーザの入力と索引ファイルを照合して該当する文書を検索する．そして，検索結果を出力インタフェースによってモンゴル語で表示する．

文書ファイル群は，http://onigoo.olloo.mn から収集した500件のキリル文字新聞記事を３．で説明した翻字システムによってあらかじめ作成した．さらに，この文書ファイル群から自動的に索引ファイルを生成し，転置ファイルを構築する．

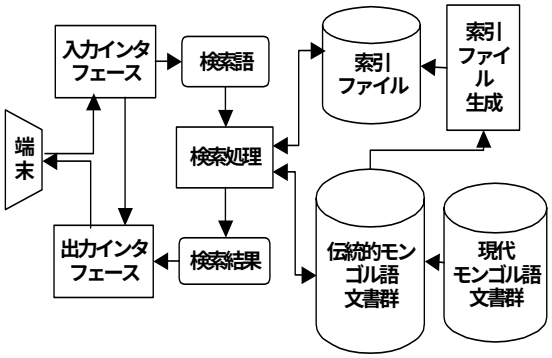


図３　モンゴル語の言語横断検索システム

6. 検索システムの評価実験

検索システムの機能評価には，事前に用意したテストコレクションが必要である．しかし，モンゴル語には，利用できる電子化されたテストコレクションが存在しない．そこで，今回は現代モンゴル語テキスト 500 件を対象に，本研究の提案した翻字手法に従い自動的に伝統的モンゴル語に翻字しテキストをテストコレクションとして利用し，結果を人間が判定した．

翻字システムの翻字結果がテキスト検索に適しているかを評価するために，検索システムの再現率に対して評価を行う．

検索質問（単語）として与えた単語が検索対象となるデータベースに含まれているものを全て結果として多く出力すれば再現率が高くなる．即ち，テキスト検索の性能を通して，提案した翻字手法を間接的に評価する．

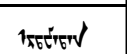
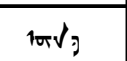
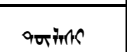
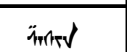
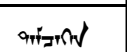
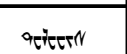
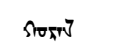
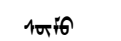
検索質問として，伝統的モンゴル語でよく使われる単語10個を用いて評価実験を行った．再現率は表１のようになった．

表１より，検索質問に適応する文書を平均で87.2%を正しく検索できたことから翻字手法が有効であることが分かった．

12.8%の検索漏れを分析した結果，翻字手法の規則を更に改善する必要があることを分かった．モンゴル文字表記で短母音を書き，長母音で読め

ることがある．例えば，モンゴル文字表記で「cagan」をキリル文字表記では「cagaan」と書くものがある．

表１　検索実験の結果

検索質問	検索記事数	検索出来た記事数	再現率(%)
	21	19	90.5
	63	51	81
	13	11	84.6
	9	7	78
	50	47	94
	20	0	100
	44	40	91
	50	39	78
	4	3	75
	10	0	100
平均	28.4	3.7	87.2

7．おわりに

本論文は，伝統的モンゴル語と現代モンゴル語のテキストを双方向に翻字する手法を提案した．特に，伝統的モンゴル語のテキスト検索システムを実現することができ，電子化テキストデータが乏しい伝統的モンゴル語のテキスト処理研究を促進することができる．

評価実験の結果，現代モンゴル語から伝統モンゴル語への翻字は 82%の精度で変換が可能であり，変換誤りがあってもテキスト内容の理解には支障のないことが分かった．このシステムを利用することによって，モンゴル国と内モンゴルの間で文書の情報交換が可能になる．

しかし，伝統的モンゴル語から現代モンゴル語の翻字の規則をさらに改善する必要がある．

参考文献

[1]Cinggeltei，現代モンゴル語文法，内蒙古人民出版社，呼和浩特市，1999.（モンゴル語）

[2] 中里致元，生出恭治．“現代モンゴル語の異種表記法の相互変換システムの構築に向けて，”情報処理学会研究報告，2002-CH-53，pp.41-46，2003

[3]那任巴図，現代蒙古語，内蒙古大学出版社，呼和浩特市，1995.（モンゴル語）

[4]トゴ，モンゴル語文法概要，内蒙古少年児童出版社，呼和浩特市，1986.（モンゴル語）

[5] 満都拉，藤井敦，石川徹也“モンゴル語全文検索システムの実現，”言語処理学会第 10 回年次大会発表論文集，pp. 129-132，2004

[6]Галсанпунцаг，Монгол улсын кирил үсгийн дүрэм（和訳：モンゴル文字とキリル文字正字法）,内蒙古人民出版社, 呼和浩特市, 2001.（モンゴル語）