多言語機械翻訳を目指したモンゴル語テキスト文法分析

伊•達瓦[1]　呼斯勤[2]、六月[2]，岳耀明[2]　吾浪[3]、白双成[2]　巴特塞恒[2]　井佐原均[1]

[1]情報通信研究機構(NICT)　日本、[2]MENKsoft CO,.LTD 中国、[3]新疆人民放送局蒙古語編輯部 中国

{idawa,isahara}@nict.go.jp

1.序論

モンゴル語はアルタイ語族に属し、語順や単語の構造及び日本語の「て/に/を/は」にあたる語尾助詞や動詞活用形などは日本語とよく似ている。日本語の文を作る感覚で単語を並べて行けばモンゴル語 - 日本語（　逆にしても良い）の文の翻訳やモンゴル語諸テキストへの翻字が可能となる。

機械翻訳に用いるモンゴル語テキストは図１に示したように、使用する国や地域によって、異なる文字システムで書かれ、異なる処理システムを使って処理されている。使用者によって縦書きのものと横書きのものがあり、地方方言によって単語や文の構造も異なる場合が多い。したがって、本研究では、それらのテキスト間との通信や変換及び多言語への翻訳を目指して、まずモンゴル語諸テキストをローマ字（Latin）表記の中間テキストに書き換え，モンゴル語のコーパスとして利用している。本論文では現代モンゴル語 NM(New　Mongolian，主に、モンゴル国やロシアカリミク、ブイラト連邦国使用)のテキストを処理対象として、自立語語尾に接続される助詞の分割ルール及び伝統的モンゴル語 TM（Traditional Mongolian、主に、中国内蒙古地域使用)、モンゴル語 Todo(中国新疆地域、ロシア国カリミク地域に居住蒙古民が使用)テキストへの翻字における長母音変換ルールについて検討した結果を報告する。

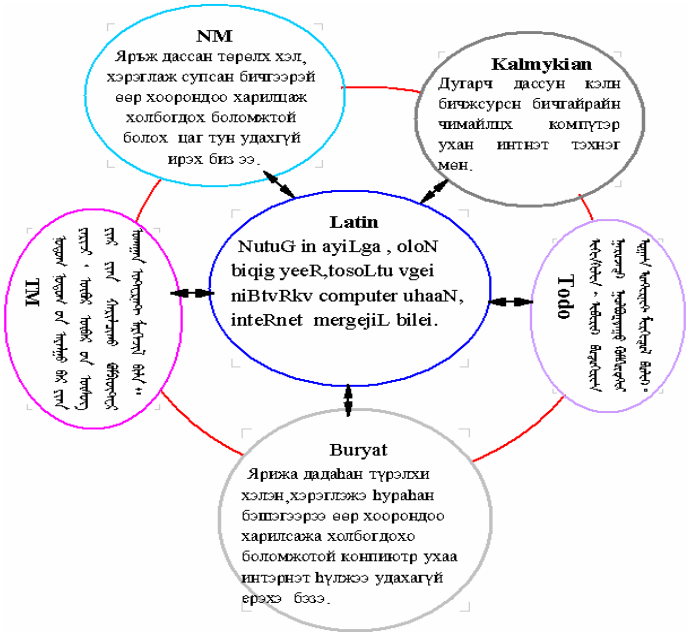

図１モンゴル語諸言語電子化テキスト例

2.モンゴル語の構造と文法

2.1 モンゴル語諸テキストの構造特徴

日本語や韓国語などのアルタイ語族諸言語に見られる「主語、目的語述語」という語順がモンゴル語諸テキストにも当てはまり、日本語と英語での語の対応に見られるような語順の入れ替えが、日本語とモンゴル語ではほとんど起こらない。また、モンゴル語の諸テキストは(図２のように)単語ごとに空白によって分かち書きができ、特に、単語の分割は不要であるが、現代モンゴル語の場合は助詞や格助詞などは直前の単語に接続して書く場合が一般的である(図２の実線〇部分)。単語から助詞を分割しておくと、NM から他の言語への翻訳や翻字などの処理には都合がよくなる。特に、TM や Todo テキストへの変換には、これが一つの不可欠な処理であり、研究[1]でも検討されている。

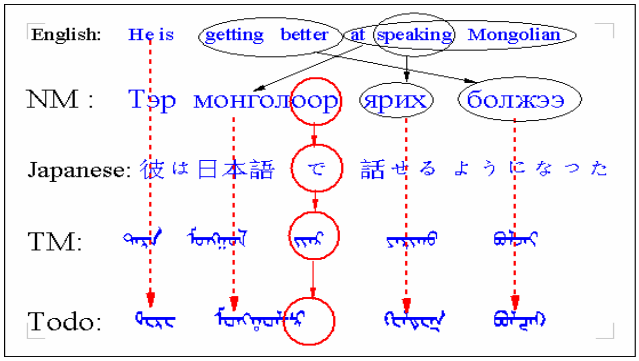

図２モンゴル語諸テキスト文法構造と日本語構造比較

また、分割された部分文字列は直前単語の母音調和によって、例えば「~aar/~oor/~eer/~o`o`r」(日本語格助詞の「で」にあたる)といった４つの交替形を持ち、また NM から TM や Todo での助詞翻字の場合は、語末文字が母音か子音かによってその表記が異なる。この点は日本語と違う[2]。例えば、NM の文字列「mongol**oor**」中の「**~oor**」はそれにあたり、(モンゴル語で)という意味を現しているが、直前の自立語「mongol」の第一音節に陽性母音「o」があるため、母音調和によって「~oor」と書けなければならない。

言語学的及びテキストデータの統計的考察による NM テキスト自立語に後続する格助詞の分割一変換規則を表１にまとめた[3]。実テキストには本規則に従い、正確に分割されるものもあれば、誤分割されるものも多い。例えば、NM テキストから空白を区切りとして抽出されたある文字列(以後形式単語と呼び、ローマ字表記する)「mongol**oor**」について、表１のルールを用いて分割すると、前例のように、 正確な意味を保って分割されるが、同じく NM 文字列「0doo zvgeer」(今大丈夫)中の形式単語「zvgeer」(大丈夫)は「zvg_eer」「方向」でまたは「矢」をとして分割され、元の文字列が正確な意味を失ってしまう。また、NM テキストの「ter bawgaaas aij zvrhn hagaran vhjee」(彼は熊に怖がらされて心臓が止まった)を、一対一の文字単位で Todo テキストへ翻字した場合は、「tere babugagaasan aiji zvrkvni hagarad vkvji」(彼は女房に怖がらされて心臓が止まった)という怪しい意味になってしまう。即ち、言語学的なルールまたは文字単位での翻字だけではかなり限界があると思われる。本研究では、多言語自立語対応辞書[4]、言語的ルール集+ DP マッチング手法(2.2 節で述べる)及び文字単位変換ーコーパス照合修正といった総合的な翻字手法を提案する。

表1:NM形式単語語尾格助詞分割一変更ルール集(表中欄NM final of a word　はNM自立語の語尾文字を、NM suffixは形式単語の語尾文字列を現す)

NM final of a word	NM suffix	TM CP	Todo CP	Jp CP	NM final of a word	NM suffix	TM CP	Todo CP	Jp CP
V_iV_jV	н	/yin/	/giin/ /in/	の	V	аар ээр оор өөр	/bar/ /ber/	/ber/	で / を
V	ийн	/yin/	/in/		C		/iyar/	/yer/ /aar/、/oor/ 1 - 2	
C	хээ	/un/ /vn/ ,/hi/	/giin/ -		V/VV +г		/iyar/	/ggar//ger/ g' aar/ -	
н	ий	/u/ /v/	/ni/		V	аа ээ оо өө хан	/ben/ /ban/	/been/ /een/-	を / に
	ы		/ni/		C		/iyen/	/yeen/ /een/-	
	хөө	/yi/	/neen/		VV +г		/iyen/	/gaan/ - g' aan/-	
V	ын	/yin/	/in/		V	аас ээс оос хан	/eqe/	/eec - e/ - -	～ から， より
C	ын	/un/ /vn/	/iin/ -		C		/eqe/		
V	ийг	/yi/	/igi/	を / のを	V C	тай той тэй таа	/tai/ /tei/	/tai/- /tei/-	と
C	ийг	/i/ ?	/iigi/ -						
V	ыг	/yi/	/igi/		V C	даа доо	/dagan/	/daan/ -	に / へ
C	ыг	/i/ ?	/igi/						
н(ng) V_iV_j	ыг ийг	/yi/	/gi/		V C	дээ дөө	/degen/	/deen/ -	
б/г/р с/д/	т	/tu/ /tv/	 	に	V C	таа тоо	/tagan/	/taan/ -	
V/ном/д /н(ng)/	д	/du/ /dv/	 		V C	тээ төө	/tegen/	/teen/ -	
other	нд								
н(ng)	ууд	/uu/	1.	～ たち ～ら	V C	гүй гэй	/vgei/	/vgei/ /go/-	ない
Other C	үүд	/vv/d/	2.						

NM: New Mongolian; TM: Traditional Mongolian; Todo: Mongolian Todo;V: Vowel; C: Consonant; 1. Positive, 2.Negative; / ～/ : reading; " - " : link to pre-word;CP:Case particle; i≠j

2.2 DPマッチングによる単語類似度計算

前述した各文構造特徴によって実際に用いているテキスト形式単語或は分割された部分文字列を自立語辞書から直接調べると、マッチングできない場合が多い。本研究ではDP(Dynamic Programming)マッチングを適用することによって、高速高類似度単語照合を行う[5]。DPマッチングのアルゴリズムは、次のように定義される。例えば、形式単語や辞書単語それぞれをA,Bとし、式(1)で与えられる。(a_i, b_j それぞれはA,B中のi, j 番目の文字を表す)

$$\begin{cases} A = a_1, a_2, \cdots a_I \\ B = b_1, b_2, \cdots b_J \end{cases} \quad (1)$$

(I,Jは単語長となる)。

次に、文字同士間距離を$d_n(i,j)$として、式(2)で計算する。

$$d_n = \begin{cases} 1 & if \quad a_i = b_j \\ 0 & else \end{cases} \quad (2),$$

$$g_n[0][0] = \begin{cases} 0 & if \quad A[0] = B[0] \\ 1 & else \end{cases} \quad (3)$$

$$For \quad j = 1 \quad to \quad |B|, \quad g_n[0][j] = g_n[0][j-1] + d_n[0][j]$$

$$For \quad i = 1 \quad to \quad |A|, \quad g_n[i][0] = g_n[i-1][0] + d_n[i][0] \quad (4)$$

そこで、単語間正規化最小値関数は(初期値は式(3),(4)とする)下式(5)として定義される。(n は辞書長を示す)

$$g_n(i,j) = \min \begin{cases} g_n(i, j-1) + d_n(i,j) \\ g_n(i-1, j-1) + 2 \times d_n(i,j) \\ g_n(i-1, j) + d_n(i,j) \end{cases} \quad (5)$$

したがって、k 個の候補単語距離を$D^k_{\min} = (1/I+J) g_k(I,J)$とすると、最終的に選択される単語距離は、条件($C^k_{\max_slope} \geq I/2$によって,下の式(6) として決定される。

$$D_x = \min\{ w_k \times D^k_{\min}(A,B) \} \quad (6)$$

ここで、重み$w_k = c^k_{\max_slope} / I$ は、形式単語長I 及びk 番目の候補単語が$i = j$ で連続してマッチした部分文字列最大値$c^k_{\max_slope}$を用いて計算される。

2.3 NM文字列後続助詞分割一変換

分かち書きNMテキストから空白によって抽出されたある文字列、すなわち、ある形式単語に対して、次の三つのス

テップで分割-翻字処理が行われる。

Step1: ある形式単語に対して、モンゴル語多言語自立語対応電子化辞書 MPEDMCJK（セクション３で紹介する）を用いて， まず、該当文字列は自立語として存在するかどうかを調べる。次に、もし、調べた結果が“ yes ”なら、該当単語に対応した他言語自立語がそのまま出力される。

Step2: 調べた結果が“ No ”なら、表 1 を利用する。そこで、もし、該当する形式単語語尾が表１のルールに従うものであれば、マッチングされた語尾は一旦切り出され、残りの文字列に対して、NM 自立語辞書との DP マッチングが行われ、高類似度の k 個の候補単語が選択される。そこで、式(6)を用いて最小距離となる候補が選択され、それに対応した他言語単語が出力される。一方、一旦切り出された語尾は表１に定義された他言語対応格助詞(CP)へ入れ替え、前単語と空白を空けて出力される。例えば、NM 形式単語を、“ aawd ”とすると、Step1 では、本文字列が見つからない。そこで、Step2 の処理を行うと、“ aawd ”は“ aaw_d ”に書き換え、また、前半部分の文字列“ aaw ”に対して NM 辞書との DP マッチングを行うと自立語「aaw」が選択され、それに対応した TM 単語「abu」が出力される。同時に、単語「abu」の語尾は母音であることから、表１を参照して語尾の「d」は「du」に置き換えられる。

Step3: Step1,2 では変換できなかった形式単語においては，日本語への翻訳の場合は該当文字列は未知語として処理される。TM や Todo への翻字の場合は、次節に述べる長母音変換処理を先に行った後で翻字が行われる。

2.4 モンゴル語諸テキスト長母音変換処理

NM や Todo テキストでは、長母音を表記する専用の音節が存在するが、TM 文語では長母音表記音節がない。即ち、諸文語テキスト同士間の変換には長母音翻字は必須となる。

そこで、本研究では、各テキストに対して言語学的ルール考察を行った。NM から TM または Todo への長母音変換ルールを表 2,3 にまとめた。TM 文語では、非規則単語表記(習慣的な表記)が多く見られ、長母音音節を表現しない場合がある。また、NM 自立語に後続する助詞は自立語の性(陽、陰性）によって変わることもある。したがって、NM から TM や Todo への長母音音節変換は必ずしも表 2 のルール集に従うとは言えない。本研究では、表 2,3 のルール集、言語コーパス及び DP マッチング手法を同時に参照した処理手法を検討した。長母音変換処理は前節 2.3 の処理では変換できなかった形式単語において単語ごとに行う。

まず、文節の中に長母音が包含されるかどうかを調べ、もし、包含されるのであれば、文字単位変換テーブル(NM の一文字が TM や Todo の一文字に対応した表、本文中では省略する)と表 2,3 を用いて、長母音出現位置(Initial)か(medial)かによって他言語文字列へ翻字される。

次に、前処理で翻字された文字列や長母音を包含しない形式単語に対して、翻字先言語大規模テキストから作成した独立単語コーパスとの DP マッチングを行い、類似度が最も高い文字列を最終的な結果として出力する。

2.5 多義語処理対策

日本語からモンゴル語への翻訳の場合は、日本語の格助詞「~から」「~より」などは、NM の場合は、母音調和によって「~aas，~ees，~oos，~o`o`s」の４つの形式に対応するが、TM では、ただ「eqe」、また Todo テキストでは「eece」

表 2NM から TM または Todo テキスト長母音条件付け変換ルール集

long vowel(Initial) in NM		long vowel(Initial) in TM	long vowel(Initial) in Todo
аа /aa/		/aga/ , /a/	/aa/
ээ /ee/		/ege/ , /e/	ee/
оо /oo/		/ogu/	/oo/
уу /uu/		/ugu/ , /agu/	/uu/
ий /ii/		/ii/	/ii/
өө /o'o'/		/ o'ge/	/o'o'/
үү /vv/		/egv/ , /vgv/	/vv/
Eve alphabet	longvowel(medial) in NM	longvowel(medial) in TM	long vowel(medial) in Todo
ж/з р/ш	аа	iga/ , /iya/ , /aga/	/aa/
other		/iga/ , /iya/ , /vga/	/aa/ , /ia/
д(л)	ээ	/ege/ , /ebe/	/ee/ /ei/
р/ш		/iye/ , /ige/	
other		/ege/ , /ige/ , /o'ge/ , /iye/	
с	ий	/igi/	/ii/
р/ш	оо	/oga/ , /iga/ , /oo/ , /ogu/	/oo/
other		/oga/ , /iga/ , /ogv/ , /aga/	
с	уу	/agu/ /ugu/ , /uu/	/uu/
д	өө	/o'ge/ , /o'gv/ /ege/	/o’ o’ /
с		/o'ge/ , /o'gu/ , /ege/	
р/ч/ш	үү	/egv/ , /igv/ , /vgv/	vv/
ж/з		/egv/	
other		/egv/ , /igv/ , /vgv/	

表３　NM から TM または Todo テキスト無条件変換ルール集

直接変更 NM	TM	Todo
яа	/yaga/	/yaa/ ,
еэ	/yage/	/yee/ - -
юу	/yagu/	/yuu/ -
юү	/yagv/	/yvv/
ее	/yoge/	/yoo/ - /yee/-

に対応する。また、NM の一つの固有名詞が TM の複数の固有名詞(同音異義語)に対応する場合もある。例えば、図３の例のように、NM 固有名詞の「/sara/」(日月の「月」、天体の「月」)は TM 固有名詞の「/sara/(ゲツ)又は/sar_a/(ツキ)」に対応する。今後はこのような語に対する翻字において、文字コード(Unicode)の差違を用い、大規模なテキストデータから作成された同音異義語コンテキストコーパスを利用した処理方法を検討したい。

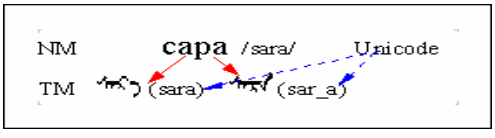

図３TM テキストに見られる同音異義語と Unicode 差違例

3.多言語自立語対応一品詞付け電子化辞書

モンゴル語自然言語処理用大規模コーパスの構築及び多言語機械翻訳システムの実現を目指して、日本独立行政法人情報通信研究機構自然言語グループと中国内蒙古自治区社会科学院モンゴル語情報処理センター(MIT),内蒙古MENKsoftCO,.LTD 社及び韓国 KAIST コンピュータ情報処理所は 2005 年から共同研究プロジェクトを実施し、この一年中は現代モンゴル語自立語５万個を軸とした“多言語対応品詞付け電子化辞書 MPEMCJK”を構築した。図 4,5 それぞれは MPEDMCJK 構造及びフォーマット設計を示す。本研究で構築する MPEMCJK 辞書は 2006 年４月に最終的な評価を行い、６月から公開される予定である[4]。

図４　MPEDMCJK 構造

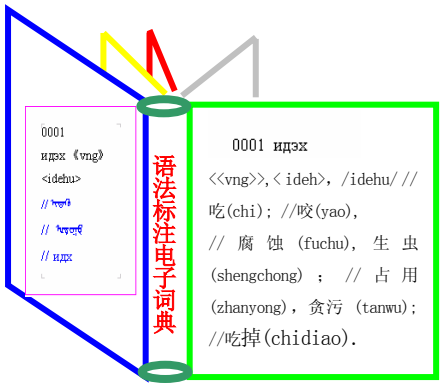

図５ MPEDMCJK フォーマット設計

4.評価

4.1 実験１．NM テキスト単語後続助詞分割実験

実験１では、MENKsoft 社提供の現代モンゴル語テキスト(19960 語)を利用した。実験は表４に示すように、３つの段階に分けて行った。表４の分割単語数はテキスト単語語尾が表１の(NM suffix)各項目に存在した形式単語数を,D(有、なし)は MPEDMCJK 辞書を利用した又は利用しない場合を表す。今回の実験評価は本手法によって処理された結果を、MENKsoft 社及び MIT 本プロジェクトメンバーが人手でチェックすることによって行った。

実験表４ NM テキスト分割実験語数 19960(*MPEDMCJK)

	D*なし DP 適用なし	D* 有 DP 適用なし	D* 有 DP 適用
分割単語数	8972	5211	5211
正解率%	64.3	82.6	91.3

4.2 実験 2. NM テキストーTM または Todo テキスト変換実験

NM->TM, Todo 変換実験は、上述実験１では変換が行われなかった全形式単語において実行された。DP マッチング言語情報コーパスとしては、本研究で作成した TM テキストから作成した 28000 語彙、Todo から作成した 34 1 27 語彙を用いた。データは MENKsoft 社及び MIT が提供した。評価を表５に示した。

表５　NM->TM または Todo 変換実験(DP　マッチング適用)

NM 単語数	TM	Todo
8190　->	7289	7486
正解率(%)	89	91.4

4.3 実験結果分析：実験結果の表 4,5 では、辞書及び DP マッチングを適用しない場合と比べ、辞書、コーパス及び DP を適用した場合は高精度な分割や翻字ができることが分かる。特に Todo への変換では最も高い変換結果が得られた。それは、Todo テキストでは、助詞や長母音の表現は NM とあまり差がないためであると思われる。今回の実験では、辞書やコーパスではマッチングできなかった多義語や活用形語尾動詞は文字単位で翻字されている。

５．おわりに

本論文では、多言語情報コーパス、言語学文法ルール及び文字列間 DP マッチング最小距離類似度を用いたモンゴル語多文字システムテキスト処理変換手法を提案した。提案手法を利用した今回の実験(NM テキストから TM 及び Todo テキストへの翻字)では、TM テキストへの変換精度は 89%で、Todo テキストへの場合は 91.4%となり、ほぼ実用レベルに達していることが確認できた。

本システムはモンゴル語を使用する地域や国間での文書情報交換には大きな貢献となる一方、モンゴル語諸テキストの自動編集-修正、品詞情報付け言語コーパスの作成及びモンゴル語自然言語処理、機械翻訳システムの実現に有用と期待される。

モンゴル語の任意のテキスト間の変換、多義語処理対策、実用化システムの実装及び効果的な評価手段の検討などは今後の研究課題となる。

参考文献

[1]満　都拉、石川　徹也ら、“伝統的モンゴル語と現代モンゴル語の双方向的な翻字”、言語処理学会第１１回年次大会　発表論文集 pp360-363(2005.3).

[2] Galcan funcag “Mongol ulassin kiril vsvgiin dvrem”,(現代モンゴル語文法)、中国内蒙古人民出版社 2001,Hohhot.（蒙古語）

[3]Erden Bagan, Gonqig sorung, “蒙古語文法”,中国内蒙古教育出版社 1998.5

[4] 伊.達瓦、井佐原　均ら、“モンゴル語言語―文字自動処理”、中国語情報処理学会誌、2005.11 採録論文 4232.

[5]Masek, W.,and Paterson, M “A faster algorithm computing string edit distance”　J. Comput. System　Sci. 20(1980), 18-31.