日本語対訳辞書拡張のための ウイグル語からウズベク語への翻字手法

小川 泰弘† 福田 ムフタル‡ 外山 勝彦†
†名古屋大学 大学院情報科学研究科 ‡名古屋産業大学 環境情報ビジネス学部
yasuhiro@is.nagoya-u.ac.jp

1 はじめに

我々はこれまでに，日本語–ウイグル語翻訳システムの開発を進めてきた [1][2]．日本語とウイグル語は，言語学においてはともに膠着語に分類され，また語順がほぼ同じであるなどの点で構文的類似性が高い．そのため両言語間の機械翻訳においては，形態素解析の結果を逐語訳することによって，ある程度の翻訳が可能である．そこで我々は，日本語の膠着語としての性質を捉えている派生文法 [3] に着目し，派生文法に基づく形態素解析システム MAJO[4] の辞書を，日本語–ウイグル語対訳辞書に置き換えることによって，日本語–ウイグル語機械翻訳システムを実現した．

ここで，派生文法は日本語・ウイグル語以外の膠着語にも適用可能であり，MAJO の辞書を置き換えることにより，他の膠着語への機械翻訳にも応用できる．そうした観点から，現在，我々は日本語–ウズベク語翻訳への応用を進めている．

ウイグル語とウズベク語は共にアルタイ語族チュルク諸語南東語群に属する語である．また，ウイグル語が使用される新疆ウイグル自治区とウズベク語が使用されるウズベキスタン共和国は地理的にも近く，両言語は文法的に類似するだけでなく，語彙的にも類似点が多い．そのため，両言語の話者間では，通訳なしで意志疎通が可能である．

よって，我々が開発してきた日本語–ウイグル語機械翻訳システムを利用すれば，少ないコストで日本語–ウズベク語機械翻訳システムが実現可能と考えられる．特に，MAJO の辞書を日本語–ウズベク語辞書に変更すれば，基本的なシステムが実現できると考えられる．

しかし，電子化された日本語–ウズベク語辞書はいまだ存在しない．印刷物としても，ウズベキスタン国内での学習者向けに小規模なものが存在するだけである．一方，ウズベク語–日本語辞書に関しては，文献 [5] が出版されている．そこで我々は，文献 [5] を元に日本語–ウズベク語電子化辞書の作成を進めている．しかし，登録されている日本語見出し語数は，現在のところ約 1.8 万語であり，語彙数の不足は否めない．

そこで，実用的な翻訳システムの実現には，対訳辞書の拡張が必要となる．対訳辞書の拡張に関しては，言い換えを利用した方法や，他の言語との対訳辞書 (主に英語) を利用する方法などが提案されている．

しかし，ウイグル語とウズベク語の間の類似性に着目すれば，別の手法での対訳辞書の拡張が可能となる．すなわち，ある日本語単語の訳語が，日本語–ウイグル語辞書には含まれているが，日本語–ウズベク語辞書に無い場合，ウイグル語訳語をウズベク語訳語に変換して，日本語–ウズベク語辞書を拡張する方法である．

ここで，そうしたウイグル語からウズベク語への変換は，両言語の類似性に着目すれば，翻訳というよりは翻字として捉えることができる．類似した言語間の翻字に関する研究である文献 [6] では，伝統的モンゴル語と現代モンゴル語との間の変換規則を人手によって作成している．一方，類似した言語間でなくとも，固有名詞などの翻訳では翻字が必要になり，文献 [7] では，統計的手法により翻字を試みている．

そこで，本研究では，ウイグル語からウズベク語への翻字手法について，人手で作成した規則による方法と，統計的な手法の両者を試みて，その実用性について議論する．

なお，本稿ではウイグル語はローマン体，ウズベク語はイタリック体で表記する．また，対応する単語同士を〈 〉で囲んだ 3 項組，または 2 項組で表現する．

2 ウイグル語とウズベク語

表 1 にウイグル語とウズベク語の対応を示す．なお，最右列の類似度については 3.2 節で述べる．表 1 において，bulut, fontan, yiring の 3 語は，まったく同じ形であり，それ以外の語も良く似ている．このように，両言語間には語彙的な類似点が多い．しかし，当然，相違点もあり，それらは以下の三つに分類できる．

表 1: ウイグル語とウズベク語の対応

日本語	ウイグル語	ウズベク語	類似度
雲	bulut	*bulut*	1
噴水	fontan	*fontan*	1
うみ (膿)	yiring	*yiring*	1
頭	bax	*bosh*	2
恐怖	déḫxét	*dahshat*	2
雇う	yallimaq	*yollamoq*	2
育てる	östürmék	*ochirmoq*	2
木	déréh	*daraxt*	3
タンク	tanka	*tank*	3
分数	késir	*kasr*	3
電気	elektir	*elektr*	3
狙撃兵	atqquqi	*snayper*	–

2.1 表記方法の相違

まず，発音は同じだが，表記が異なる場合がある．例えば，表 1 の〈恐怖, déḫxét, *dahshat*〉や〈頭, bax, *bosh*〉において，ウイグル語で x と表記される箇所と，ウズベク語で *sh* と表記される箇所の発音は同じである．

2.2 発音の相違

次に，表記だけではなく，発音そのものが異なる場合である．特に，ウイグル語には母音調和の規則があるが，ウズベク語にはない点が大きな差異である．

母音調和とはアルタイ諸語などに見られる現象で，単語に現れる母音の組合せが一定になることである．例えば，表 1 のウイグル語〈雇う,yallimaq〉の接尾辞 -maq と〈育てる,östürmék〉の接尾辞 -mék は同じ機能をもつが，母音調和により形が異なっている．一方，ウズベク語には母音調和の規則がないため，*yollamoq* と *ochirmoq* のように，同じ形の接尾辞-*moq* が接続する．

2.3 単語の由来の相違

最後に，単語の由来の違いによる相違がある．例えば，表 1 の〈狙撃兵,atqquqi,*snayper*〉において，ウズベク語の *snayper* は英語の sniper と同じ語源をもつ語であるが，ウイグル語の atqquqi は動詞〈撃つ,atmaq〉から派生した語である．

3 対応する単語ペアの作成

表 1 に示したウイグル語とウズベク語の対応を元にウイグル語からウズベク語への翻字を考える．表 1 は人手により作成したものであるが，本章では，そのような対応のペアを自動的に収集する手法について述べる．

本研究では，日本語–ウズベク語電子化辞書と日本語–ウイグル語電子化辞書を利用し，日本語を中間言語と考えることにより，類似するウイグル語とウズベク語のペアを，以下の手順で作成した．

3.1 対訳辞書からの抽出

今回使用した日本語–ウイグル語電子化辞書は 36,157 語の日本語見出しを含む．一方，日本語–ウズベク語電子化辞書は 18,526 語の見出し語を持つ．

これらの辞書から，同じ日本語見出し語をもつ辞書項目のペアを 3 項組として抽出した．なお，日本語見出し語に複数の訳語がある場合には，それらすべてを抽出しているため，3 項組の内容は〈日本語見出し語, ウイグル語訳の集合, ウズベク語訳の集合〉である．例えば，日本語見出し語「頭」に対して，日本語–ウイグル語辞書には 8 語，日本語–ウズベク語辞書には 3 語の訳語がそれぞれ掲載されていたため,〈頭, {éqil, bax, baxliq, baxqi, kalla, kattiwax, mingé, uq}, {*bosh, kalla, katta*}〉という 3 項組となった．そのような 3 項組は 5,580 組得られた．

3.2 類似する 3 項組の抽出

3.1 節で抽出された 3 項組には，複数のウイグル語とウズベク語が含まれるが，翻字の参考になるのは，ウイグル語とウズベク語が類似するペアだけである．前述の「頭」の例では,〈頭, bax, *bosh*〉や,〈頭, kalla, *kalla*〉の二つの 3 項組のみが有用である．一方,〈狙撃兵, atqquqi, *snayper*〉のように，日本語訳は同じでも，単語同士が類似しないものは除きたい．そこで，ウイグル語単語とウズベク語単語間の類似度を以下の 3 段階 (類似していない場合を含めれば 4 段階) に分類した．

まず，**類似度 1** となるのは,〈雲, bulut, *bulut*〉のように単語の綴が完全に一致する場合である．

類似度 2 となるのは，以下の変換によってウイグル語単語とウズベク語単語が一致した場合である．

1. ウズベク語中の sh を x に,ch を q に,x を h に変換.
2. 子音が後続する ng を n に変換.
3. 以下の文字を対応する文字にそれぞれ変換.

   i, e, a, o, u, é, ö, ü, y, o‘ → i
   k, q, g, ķ, ğ, g‘→ k
   f, p, b, w, v → f
   t, d → t
   s, z → s
   n, m → n

表 2: ウイグル語–ウズベク語ペアの抽出結果

類似度	3 項組	2 項組
1	192	161
2	3,023	2,281
3	360	218
合計	3,575	2,660

4. ウズベク語単語中に残った ‘を除去.
5. 同じ文字の連続を一つにまとめる.

なお，この変換は Soundex[8] を参考にした．類似度として編集距離を用いることも可能であるが，単純な編集距離では，ü が u になるような起こりやすい変化と，ü が k になるような起きにくい変化が同じ距離になるため，上記の変換を採用した．この結果，例えば bax と *bosh* はともに fix に変換され，類似度 2 となる．

ただし，この変換では，〈木, déréh, *daraxt*〉,〈タンク, tanka, *tank*〉のように末尾の音素が消失する場合や,〈分数, késir, *kasr*〉,〈電気, elektir, *elektr*〉のように子音に挟まれた母音が消失する場合に対応できない．

そこで，上記の変換アルゴリズムの適用後，さらに

(1) 一方の末尾を削除したものが，もう一方と一致.
(2) 一方の末尾が子音・母音・子音であり，その母音を削除したものが，もう一方と一致.

のいずれかの条件を満たしたものを**類似度 3** とした．

表 1 の最右列は，このように計算した類似度である．なお,〈狙撃兵, atḳuqi, *snayper*〉の場合は，いずれの方法でも一致しないため，類似していないと判断される．

この類似度による判定を用いた結果，ウイグル語とウズベク語が類似している 3 項組を 3,575 組獲得した．類似度別の内訳を表 2 の「3 項組」の欄に示す．

3.3 類似する 2 項組の抽出

前節で得られた結果には,〈頭, bax, *bosh*〉だけでなく,〈頭目, bax, *bosh*〉,〈先, bax, *bosh*〉のようなウイグル語とウズベク語の単語は同じだが，日本語見出し語が異なる 3 項組が含まれている．しかし，今回の翻字で必要なのはウイグル語とウズベク語の 2 項組である．そこで，日本語見出し語だけが異なるものをまとめたところ，3,391 個のペアが得られた．しかし，その中には〈味, dit, *did*〉と〈味, dit, *tot*〉のように同じウイグル語に対して，複数のウズベク語が対応しているペアが獲得された．そのような場合には類似度が小さいもの，さらに，類似度が同じ場合には編集距離が小さいものを選択した．その結果，2,660 組のウイグル語とウズベク語の 2 項組が得られた (表 2).

4 ウイグル語–ウズベク語翻字

本章では，前章で獲得した 2,660 組のペアを利用したウイグル語からウズベク語への翻字について検討する．翻字手法としては，人手で規則を作成する方法 [6] と，統計的な手法 [7] によるものが考えられる．以下，それぞれの手法について検討する．

4.1 人手で作成した規則による翻字

まず，翻字規則を人手によって作成する手法について述べる．ベースラインは，ウイグル語の各文字を対応するウズベク語の文字に 1 文字ずつ置き換える手法である．すなわち，表 3 の上の行に示すウイグル語文字を，それぞれ下の行に示すウズベク文字に置換する規則である．この規則だけを使用して，2 項組 2,660 組に含まれるウイグル語単語を変換したところ，正しいウズベク語単語に変換できたのは 992 個，すなわち変換精度は 37.3%であった．

このベースラインとなる規則に，人手による規則を追加して変換精度の向上を図った．例えば，ウイグル語の動詞接尾辞 -mék は表 3 の規則では-*mek* となるが，表 1 における〈育てる, östürmék, *ochirmoq*〉の例に示す通り，これは -*moq* となるのが正しい．そこで，単語末尾の mék は *moq* になるという規則を追加した．このように，変換規則は正規表現にマッチした部分を置換する形式で作成した．そのような規則を 21 個追加したところ，1,190 個 (44.7%) が正しく変換できた．さらに規則を追加して変換精度を上げることは可能であるが，現在の規則に新たな規則を追加しても，正しく変換できるようになるのは数個がせいぜいであり，人手によるこれ以上の精度向上は容易ではない．

4.2 統計的手法による翻字

次に，統計的な手法を利用した翻字について述べる．統計的な翻字は，各文字を単語と見做せば，統計的翻訳として捉えることができる．そこで，統計的翻訳用に公開されている各種のツールを使用してウイグル語からウズベク語への統計的翻字を試みた．具体的には，翻訳モデルの学習に GIZA++[9]，言語モデルの学習に SRILM[11]，デコーダに Moses[10] を使用した．

翻訳モデルの学習には，前章で得られた 2 項組 2,660 組を使用した．言語モデルの学習には，日本語–ウズベク語辞書に含まれるウイグル語訳語を使用し,〈喪服,*aza kiyimi*〉のようにウズベク語訳語が 2 語以上から成る場合は，1 語ずつに分割した．こうして抽出したウズベク語 23,494 語を言語モデルの学習に使用した．

表 3: ウイグル語からウズベク語への変換規則 (ベースライン)

ウイグル語文字	a	b	d	é	e	f	ğ	g	ḫ	h	i	j	ķ	k	l	m	n	ö	o	p	q	r	s	t	ü	u	w	x	y	z
ウズベク語文字	*o*	*b*	*d*	*a*	*e*	*f*	*g‘*	*g*	*h*	*x*	*i*	*j*	*q*	*k*	*l*	*m*	*n*	*o‘*	*o‘*	*f*	*ch*	*r*	*s*	*t*	*u*	*u*	*v*	*sh*	*y*	*z*

表 4: 上位 10 位までの変換精度 (累積)

順位	1	2	3	4	5	6	7	8	9	10
正解数	2,071	2,111	2,133	2,149	2,182	2,196	2,213	2,226	2,243	2,253
変換精度	77.9	79.4	80.2	80.8	82.0	82.6	83.2	83.7	84.3	84.7

学習に使用したデータをすべてテスト・データとしたクローズド・テストでは，変換精度は 77.9%であった．また，Moses では，複数の結果を順位付きで出力可能なことから，Moses に上位 10 語までの候補を出力させた．その中に正解が含まれる割合を表 4 に示す．なお，上位 10 語の出力といっても，入力となる単語が短いため，同じ出力が含まれている場合もある．例えば,〈傾ける, égmék〉に対する出力の上位 10 語はすべて *egmoq* であった．上位 10 語の出力の平均異なり数は 1.6 語であり，ほとんどが同じ単語を出力していた．

翻訳モデルの学習用データを分割し，10 分割交差検定を試みた場合の変換精度は 73.0%であり，この場合も人手による規則作成よりも良い結果が得られた．今回の実験では，パラメータの調整をしていないため，その調整によっては，精度が更に向上する可能性もある．

5 対訳辞書拡張実験

次に，前章で学習した統計的翻字モデルを用いた対訳辞書拡張について検討した．例えば，日本語–ウイグル語辞書には〈ミルク,süt〉が存在するが，日本語–ウズベク語辞書には「ミルク」の項目はない．しかし，süt を翻字モデルで変換したところ，*sut* が出力され，これは日本語–ウズベク語辞書にある〈牛乳,*sut*〉と一致した．つまり，翻字によって,〈ミルク,*sut*〉という項目を日本語–ウズベク語辞書に追加可能となった．

そこで，日本語–ウイグル語辞書に含まれているが，日本語–ウズベク語辞書に無い日本語見出し語のうち，長さ 4 文字以下のものをランダムに 1,000 語抽出し，これらに対する辞書拡張を試みた．すなわち，日本語見出し語に対応するウイグル語単語を前章の翻字モデルで変換し，その出力がウズベク語辞書に存在する訳語と一致するかどうかを調べた．上位 10 個までの出力を調べたところ，795 語に対して対応するウズベク語単語が存在した．もちろん，その中には正しくない対応もあり，例えば,〈裁判,sot〉に対する翻字の結果も *sut* になったが，これは誤りである．このように，出力結果を改めて確認する必要があるが，日本語見出し語に対するウズベク語訳語の候補が出力されることから，この翻字モデルは辞書拡張として有効である．

6 おわりに

本稿では，翻字を利用した日本語–ウズベク語対訳辞書の拡張について述べた．紙面の都合により本稿では触れないが，統計的翻字モデルにおけるウズベク言語モデルの学習用，および辞書拡張の正解候補として，いずれの場合もウズベク語の単語を多く収集することが精度向上に重要な役割を果たすと予想される．そのため，今後はウズベク語で書かれたコーパスを収集する必要がある．また，今回はウイグル語辞書を使用してウズベク語辞書を拡張したが，逆方向の，ウズベク語辞書を用いたウイグル語辞書の拡張についても試みる予定である．

謝辞 本研究は，文部科学省科学研究費補助金若手研究 (B) (課題番号 18700141) の補助を受けている．

参考文献

[1] 小川, ムフタル, 杉野, 外山, 稲垣：派生文法に基づく日本語動詞句のウイグル語への翻訳, 自然言語処理, Vol. 7, No. 3, pp.57–78 (2000).
[2] ムフタル, 小川, 稲垣：日本語–ウイグル語機械翻訳のための格助詞の変換処理, 自然言語処理, Vol. 8, No. 3, pp.123–142 (2001).
[3] 清瀬: 日本語文法新論–派生文法序説–, 桜楓社 (1989).
[4] 小川, ムフタル, 外山, 稲垣：派生文法による日本語形態素解析, 情報処理学会論文誌, Vol. 40, No. 3, pp.1080–1090 (1999).
[5] 小松: ウズベク語辞典–新版–, 大阪ウズベキスタンの会 (2004).
[6] 満, 藤井, 石川：伝統的モンゴル語と現代モンゴル語を対象とした双方向的な翻字手法, 情報処理学会論文誌, Vol. 47, No. 8, pp.2733–2745 (2006).
[7] AbdulJaleel,N., Larkey,L.S.: Statistical transliteration for english-arabic cross language information retrieval, Proc. of the 12th Int. Conf. on Information and Knowledge Management, pp.139–146 (2003).
[8] Knuth,D.: Sorting And Searching, The Art Of Computer Programming, Vol. 3, Addison-Wesley (1973).
[9] Och,F.J.: GIZA++: Training of statistical translation models. http://www.fjoch.com/GIZA++.html.
[10] Moses – a factored phrase-based beam-search decoder for machine translation. http://www.statmt.org/moses/ .
[11] SRI Speech Technology and Research Laboratory: SRILM - The SRI Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm/ .