日本語-ウイグル語機械翻訳に関する評価実験

Muhtar Mahsut† 小川 泰弘‡ 杉野 花津江‡ 稲垣 康善‡
†名古屋大学大学院国際開発研究科 ‡名古屋大学大学院工学研究科

1 はじめに

日本語とウイグル語には、名詞接尾辞や動詞接尾辞の存在及び機能の類似性、文節の文中での順序の同一性など、構文的に多くの共通する特徴がある。従って、日本語からウイグル語への機械翻訳を行なう場合、形態素解析によって得られる各単語を逐語翻訳するだけで、ある程度の翻訳結果を得ることができる。しかし、従来の日本語文法では、動詞が活用することを前提にしていたため、ウイグル語への翻訳の前に、動詞の活用処理が必要であった。そこで、我々は日本語とウイグル語を共に派生文法で記述することにより、両言語間の形態論的類似点を明確にし、単純でかつ体系的な機械翻訳が可能なことを示した [2]。また、訳語選択に関わる単語の多義性、格助詞を含む名詞接尾辞が一対一に対応しないなどの問題があり、単純な逐語翻訳では不自然な翻訳となる場合があった。そこで、特に動詞句の訳語選択に関しては、単語間の接続関係を考慮した訳語置換表を用いることにより解決し、より自然な翻訳を実現した [3]。名詞接尾辞の対応付けに関しては、両言語の名詞接尾辞の機能の類似性を活かした格パターン構造を利用して翻訳精度をさらに上げることができた [1]。

一方、ウイグル語は母音調和の激しい言語の一つであり、母音調和の影響範囲が文節全体に及ぶこともよくある。従って、日本語-ウイグル語機械翻訳を行なう場合、訳語を選択し、接尾辞を正しく接尾させるだけでは自然な訳出文にならず、語幹と接尾辞、さらに接尾辞と接尾辞が隣接した時の母音の相応な調和を正しく行なう必要がある。そこで、我々はウイグル語の母音調和現象に関して詳細な調査を行ない、音韻処理ルールを作り、機械翻訳システムに導入した。

我々は、ウイグル語-日本語辞書 [4] に現れる例文の中から100文をランダムに選び、上述のアプローチに基づいて試作した日本語-ウイグル語機械翻訳システムを用いて機械翻訳の評価実験を行なった。今回の実験は、動詞接尾辞、名詞接尾辞、及び音韻変化部分を併せたシステムに関して行ない、満足のいく結果を得ることができた。本稿では、その結果について述べる。

2 日本語-ウイグル語機械翻訳システムの実現

我々は、日本語-ウイグル語機械翻訳システムを日本語形態素解析システム、訳語置換システム、ウイグル語整形システムの三つのモジュールで構成した。本システムの概要及び動作例を図1に示す。日本語形態素解析システムには、MAJO[2] の改良版を使用した。MAJO は派生文法に基づいた日本語形態素解析システムであり、その内部辞書は、(日本語単語、品詞名、意味) の3項組の形になっているが、今回評価実験を行なった機械翻訳システムでは、MAJO の辞書を (日本語単語、品詞名、ウイグル語訳) の3項組で構成される日本語-ウイグル語辞書に置き換えて使用した。このようにしてできた MAJO の出力結果は、そのままウイグル語への逐語翻訳の結果となる。訳語置換システムは、動詞接尾辞置換表と名詞接尾辞置換表を実現するモジュールである。動詞接尾辞置換表は、(日本語動詞接尾辞、ウイグル語の基本訳語、前接ウイグル語、後接ウイグル語、ウイグル語の新訳語、新訳語の (新) 品詞) の6項目からなる表である。名詞接尾辞置換表は (日本語名詞/ウイグル語訳、日本語名詞接尾辞/ウイグル語訳、日本語動詞/ウイグル語訳) の3対か

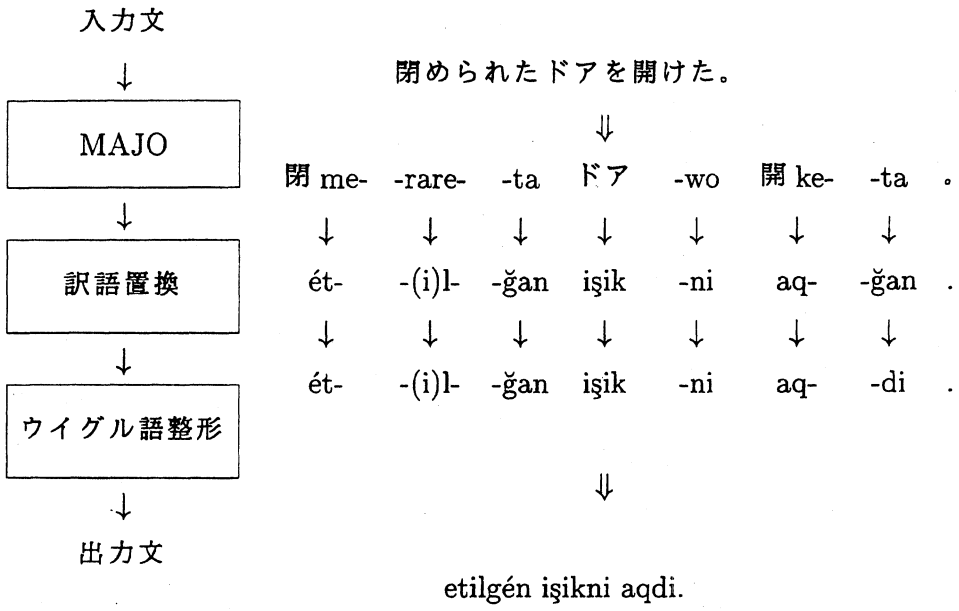

図 1 日本語-ウイグル語機械翻訳システムと動作例

らなる表である。ウイグル語整形システムはウイグル語の音韻変化規則を処理し、最終的にウイグル語訳出文を出力するシステムである。

3 翻訳実験の評価

我々は、前節で述べた日本語-ウイグル語機械翻訳システムに対して翻訳実験を行った。翻訳の日本語対象文として、ウイグル語-日本語 [4] に現れる例文の中から 100 文をランダムに選んだ。その辞書に現れる例文は、ウイグル語文とその日本語訳のペアで出現している。日本語例文を本翻訳システムに入力してそれとペアになっているウイグル語文を本翻訳システムの訳出結果との比較に用いた。これは、翻訳結果を主観的に評価するのではなく、実際の文と比較して評価するためである。本翻訳システムの辞書としては我々が作成した最新の日本語-ウイグル語電子辞書 [5] を用いた。この日本語-ウイグル語電子辞書は、見出し語が約 20,000 語の辞書である。しかし、例文に出現する単語の一部がこの辞書の見出しに存在しない場合には、それらの単語を辞書に登録した。従って、今回の実験は、未登録語はないとして行なった。また、今回の実験は、主に訳出結果であるウイグル語文の動詞接尾辞、名詞接尾辞、及び音便変化部分の翻訳精度を評価するものである。その重要な処理として、意味選択があるが、この問題は、日本語-ウイグル語機械翻訳だけではなく、機械翻訳全般における共通の問題であり、ここでは単語の「主訳」選択にとどめ、評価実験の対象項目にしていない。

表 1 翻訳実験の正解率

日本語例文数	正解翻訳数	正解率
100	81	81%

この実験では、完全に正解のものと、動詞接尾辞、名詞接尾辞、及び音便変化部分が正しく処理されたもの (人称語尾、複数形を示す語尾の間違いは許して) を正解とした。翻訳実験の正解率を表 1 に示す。すべての (意味選択処理も含む) 不正解翻訳の理由

を列挙して表2に示す。訳出文には、誤りが2箇所以上存在するものもあり、そういう意味で表2の数字には、重複するものもある。表2から分かるように、動詞接尾辞、名詞接尾辞、及び音便変化部分の翻訳精度に重点を当てた実験結果としては、満足のいくものになったと評価できる。しかし、意味処理

表2 翻訳失敗理由の内訳

翻訳失敗の理由	日本語例文数	例 (括弧内は正しい訳)
動詞接尾辞の誤り	2	建物の修理はまだ終っていない。 öy-imarétning jöndéş tehi tügéwatméydu (öy-imarétning rimonti tehi tügimidi.)
複数形接尾辞の処理	3	空に雲がゆらゆらたなびいている。 asmanda bulut gilding-gilding üzüwatidu. (asmanda bulutlar gilding-gilding üzüwatidu.)
日本語文の文節・形態素の誤り	3	私は:そう:なるとは:到底:思っていなかった。 (私は:そうなるとは:到底:思っていなかった。) mén ķéwét(層) ķurulsam zadi oylawatmadim. (mén şundaķ bolar dép zadi oylimiğan idim.)
人称接尾辞の誤り	4	彼は妻とむつまじい u ağiça bilén ép. (u ağiçasi bilén ép.)
名詞接尾辞選択の誤り	5	彼は仕事に夢中になって、食事さえ忘れてしまった。 u émgékké düm çüşüp, ğizamu untup kétti. (u émgékké düm çüşüp, ğizanimu untup kétti.)
音韻処理の誤り	10	君の話は私を引き付けた。 sening gewing meni ķarattim. (sening geping meni ķaratti.)
概念構造の違いによる失敗	12	多ければ多いほどいい。 köp bolsa köpçilik dép (köp bolsa şunçé yahşi.)
意味選択・意味処理の誤り	33	私は近い内に北京に行きます。 mén yeķin içta Beijingğa birimén. (mén yeķin(近い内)da Beijinggé barimén.)

を本格的に行なっていない分、翻訳失敗の最大の理由になっていることも分かった。次に、翻訳失敗の原因に関して考察を行なう。動詞接尾辞の誤りは、100文中に2文あり、2文とも「...している」の否定の場合である。この場合、もし、一種の「状態」を表すなら、ウイグル語の訳は、「進行中」の否定ではなく、単なるその動詞の「過去形の否定」になる。現在の機械翻訳システムでは日本語文に現れる動詞の「...している」形が「進行中」か、「状態」かの判断を機械的に行なう仕組はまだない。

ウイグル語では、日本語と同じような名詞の複数形がある。しかし、日本語で複数形接尾辞が付かないものでも、ウイグル語では付く場合がある。例えば、「空に雲がゆらゆらたなびいている」(asmanda **bulutlar** gilding-gilding üzüwatidu) の場合、日本語では「雲達」とは言わないが、ウイグル語では'**bulutlar**'(雲達) になる。日本語と違うもう一つの点は、ウイグル語では動詞にも複数形があることである。現在のシステムでは、複数形の処理に関しても人称と同じ単純な選択ルールしか行なっていないが、正確に翻訳されている。

今回の実験では、日本語文の形態素解析には、MAJO[2]の改良版を利用した。形態素解析結果の誤りによって訳語選択に失敗した文は一つ、形態素解

析が正しく行なわれても失敗する文が二つあった。100文中98文を正しく形態素に分けたというこの結果は、MAJOの解析能力の高さを示している。

人称接尾辞の誤りは、適切な人称接尾辞を文節に付加でなかったことである。日本語と違って、ウイグル語では、名詞と動詞に人称接尾辞が付くのが一般的である。人称接尾辞は、動詞に関しては動作主に依存して決まり、名詞に関しては主にその名詞がその文脈で、ある物・人の「所有・所属」関係にあるかどうかによって決まる。例えば、「彼は妻とむつまじい」(u ağiçasi bilén ép) の場合、「彼」と「妻」の間には、一種の「所属」関係があると認められるので、ウイグル語の人称語尾'si'が付くのである。現在の機械翻訳システムでは、名詞に関しては主語と名詞が「の」関係にあるかどうか、動詞に関しては、主語の人称が明確であれば、その人称、そうでなければ、三人称を選択すると言う単純な方法でしか決めていないので、数十箇所出現した人称接尾辞の内、4箇所誤った。これは、比較的よい結果と言える。人称接尾辞をさらに正確に補うには、上記のような関係を特定する必要があり、それは今後の課題である。しかし、人称が欠けていても読み手による意味的な欠落が生じないので、今回のテストでは正解とした。

名詞接尾辞 (格助詞) 選択の誤りは、我々の提案したアプローチ [1] の効果があり、数多くある名詞接尾辞の中で、失敗例は、4箇所である。

日本語-ウイグル語機械翻訳システムでは、音韻処理を行なうウイグル語整形システムにより、音韻規則ルールを一次訳出文に対して適用する処理を行なっている。今回の実験では、100文中、10文に音韻処理の誤りがあった。この結果は、音韻処理だけの観点から見れば、90%の正解率である。ウイグル語では、語幹の母音と接尾辞の母音が互いに調和し合い、その調和現象が連鎖的に文節全体に及ぶことがある。しかし、現在の音韻処理ルールでは、動詞に含まれる後母音の有無で接尾辞の音韻変化を決めるため、適切な音便変化を行なうことができない場合がある。

4 おわりに

本稿では、我々が今まで行なってきた研究成果に基づいて試作した日本語-ウイグル語機械翻訳システムに関して行なった実験結果を述べた。今回の実験では、翻訳出力であるウイグル語文の動詞接尾辞、名詞接尾辞、及び音便変化部分の翻訳精度を調べるのが主な目的であった。今までは、個別のアスペクトに対して、例えば、格助詞なら格助詞だけを対象に翻訳実験を行ない、一つの文の助詞だけが正解なら、その文を正解と勘定した。しかし、今回は、動詞接尾辞、名詞接尾辞、及び音便変化部分を同時に調査し、その一つでも間違っていれば、その訳出文を不正解とした。それでも、約80%の正解率を得ることができた。今回の実験では、訳語選択を実験対象にしていないが、正解率に大きな影響があることも分かった。今後、ここで得られた実験結果を活かしながら、システムの訳語選択を含む処理精度の向上に努めたい。

参考文献

[1] ムフタル・マフスット, 小川泰弘, 稲垣康善: 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).

[2] 小川泰弘, ムフタル・マフスット, 外山勝彦, 稲垣康善：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).

[3] 小川泰弘, ムフタル・マフスット, 杉野花津江, 外山勝彦, 稲垣康善：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.57–77 (2000).

[4] 飯沼: ウイグル語辞典, 穂高書店 (1992).

[5] ムフタル, 小川, 杉野, 稲垣: 日本語-ウイグル語辞書の自動作成と評価, 電気関係学会東海支部連合大会講演論文集, No.552(2001).