逐語訳によるウイグル語―日本語機械翻訳の研究

マヒムットジャン.ママットジャン　　岡本 紘昭
朝日大学経営学研究科

ウイグル語は中国の新疆ウイグル自治区の主体民族ウイグル民族の言語である。ウイグル語と日本語は文法的構造、語の形態的構造及び格助詞の対応など多くの面で共通の特徴があると言われている。ウイグル語と日本語のこの様な特徴を利用すれば、言語構造に関する複雑な解析をかなり回避することができて、品質の高い機械翻訳ができると予想される。

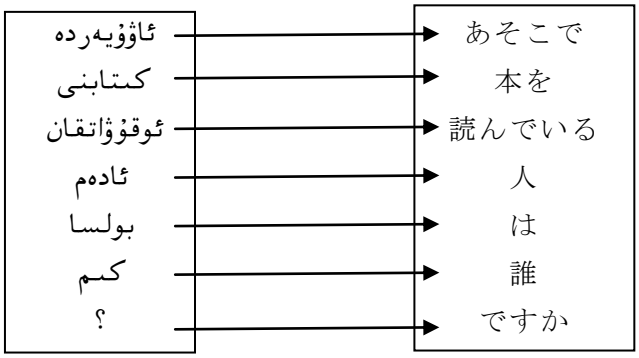
本研究では逐語訳によるウイグル語―日本語機械翻訳を試み、その精度を検討する。

1. ウイグル語と日本語の共通の構文的特徴と相違点

ウイグル語と日本語の主な共通の特徴は以下の 2 点である。

第一に、両言語は文節の文中での順序が一致する [1)]。

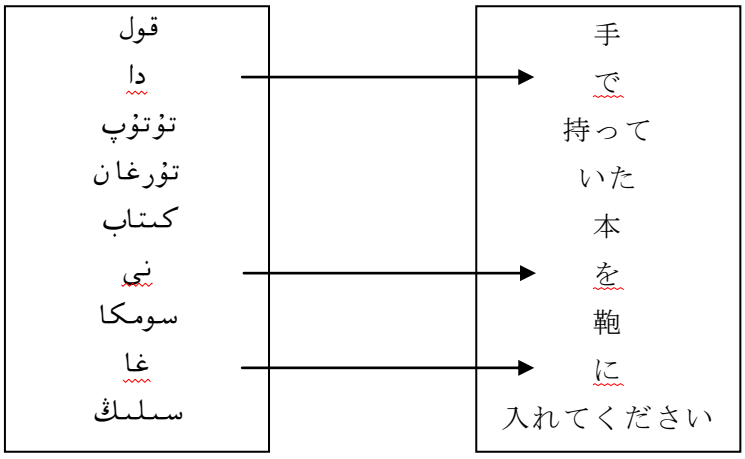
例えば：



例 1

第二に、ウイグル語と日本語は共に動詞接辞及び格助詞が存在し、その機能が類似しており、文節内部の順序も似ている [1)]。

例えば：



例 2

以上の共通の特徴に基づき図 1 に示す逐語訳プロセスを考える。

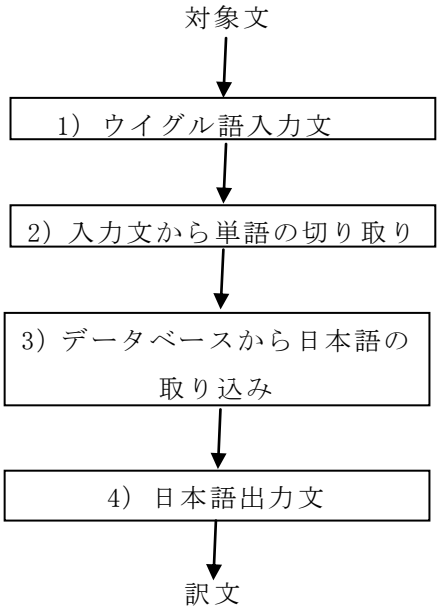


図 1　逐語訳プロセス

ウイグル語と日本語には上述の様な共通の特徴もあるが、細部を見ると色々な相違点もある。

（1）ウイグル語では日本語の「は」と「が」に対する“بولسا”がよく省略される。他にも文章の中に省略される単語、格助詞や接辞がある。

（2）ウイグル語では日本語の「です」に対応する語（単語）が無い。

（3）ウイグル語の動詞と名詞の派生型が非常に多い。

2.逐語訳プログラム

上記のプロセスに従って Visual Basic 2008 Express を利用して実験用プログラムを作った。プログラムの実行画面を図 2 に示す。

ウイグル語の正書法では、英語の文章と同じ様に単語と単語の間にスペース（空欄）を入れる。この特徴を利用して右から左に書かれたウイグル語の文章をスペースのところで単語に切り分け、このプログラムに接続した「ウイグル語―日本語機械翻訳用実験データベース」を使って対応する日本語を見つけ、取り込んだ日本語の単語を左から右に並べ変えれば、日本語訳（逐語訳）ができる。

なお、今回の実験用データベースでは、一つのウイグル語単語に対し日本語訳に複数の単語がありうる場合、最適のものをあらかじめ選んで訳語の選択の問題を回避した。

図 2　ウイグル語－日本語機械翻訳プログラム画面

3. 逐語訳プログラムの実験

このプログラムが正しく動くかどうかを確かめる為に色々な実験をしてみた。例えば、ウイグル語の文章“. مەكتەپكە بارىمەن”をこのプログラムで翻訳すると「学校へ行きます。」と正しく訳される。この文章をさらに複雑化して色々変化して見る。

1）“. مەكتەپكە باردىم”

「学校へ行きました。」

2）“. مەكتەپكە بار”

「学校へ行け。」

3）“. مەكتەپكە بېرىۋاتىمەن”

「学校へ行っています。」

4）“. ئەتە مەكتەپكە بارىمەن”

「明日学校へいきます。」

5）“. مەن ئەتە مەكتەپكە بارىمەن”

「私明日学校へ行きます。」

6）“. مەن ئەتە سەھەر مەكتەپكە بارىمەن”

「私明日朝学校へ行きます。」

7）“. ئەتە سەھەر دوستۇم بىلەن بىللە مەكتەپكە بارىمەن”

「明日朝友達と一緒に学校へ行きます。」

8）“ئەتە سەھەر دوستۇم بىلەن بىللە مەكتەپكە بېرىپ،كۈتۈپخانىدا ئۆگىنىش قىلىمەن .”

「明日朝友達と一緒に学校へ行って、図書館で勉強します。」

これ以外にも色々なパタンの 100 個以上の文章をこのシステムで翻訳して、逐語訳翻訳の特徴と問題点を概略把握し、解析を行った。その結果、殆どの訳文は正しく理解できたが、以下のようにいくつかの問題点が明らかになった。

4.逐語訳システムの問題点と解決方法について

本研究で確かめられた問題点は大きく分けて三つある

1）「は」、「が」、「です」の漏れる問題。

原文のウイグル語では省略されるために生ずる訳文中での「は」、「が」の漏れ等があった。この問題は原文ウイグル文で省略された部分を翻訳前に復元することで解決できると考える。「です」の漏れる問題は、訳文が体言で終わるときに付加すれば殆どの場合解決できる。

2）辞書データベースが大きくなる問題。

ウイグル語は派生言語であり、どんな単語も派生する。可能性として動詞は 2000 種類以上、名詞は 300 種類以上派生する[2)]。ウイグル語の中にある 40000 個の単語がごく少なく見積もって実用上 100 種類ずつ派生するとしても 400 万個になる。この数は非常に大きい為データベース作成が困難になる。本研究による逐語訳のこの欠点は、形態素解析により単語を語幹と接尾辞に分けてデータベースを作成する方法によって改善されると考える。

3）格助詞の問題

格助詞がウイグル語と日本語で一対一に対応しないため生ずる誤りがあった。ウイグル語の格助詞は元から複雑であり、関連する文法も多い。ウイグル言語学でも難しい分野の一つである。しかし、この問題も言語資料コーパスなど素材の分析及び形態素解析で改善できると考える。

参考文献

1）小川　泰弘、ムフタル　マフスット、外山　勝彦、稲垣　康善　「派生文法に基づく日本語―ウイグル語機械翻訳」　信学技報　NLC93-60(1993-12)

2）アブドレイム.アブドハリリ、伝　康晴、土屋　俊「ウイグル語接辞の頻度について」　言語処理学会第 13 回年次大会発表論文集(NLP2007)