逐語訳によるウイグル語―日本語機械翻訳の研究

Study of Word-for-word Uyghur-Japanese Machine Translation

マヒムットジャン　ママットジャン
Maihemutijiang Maimaitijiang

要旨

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。日本語もウイグル語と同じアルタイ語族に属する膠着語である。

本研究では、ウイグル語と日本語の類似点に着目し、構文解析と形態素解析を省いた逐語訳による、ウイグル語から日本語への機械翻訳を試みた。逐語訳の結果を評価するためにアンケート調査を行ったところ、通常の文章は十分理解できる程度に翻訳できることが確かめられた。また、発生する問題点の種類や頻度を調べるとともに、解決法のアイディアを提案した。

1. まえがき

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、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。

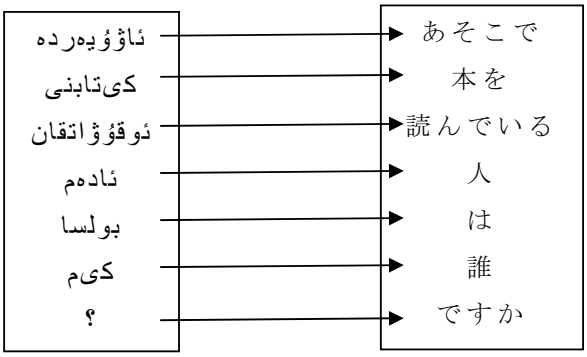
本研究では逐語訳によるウイグル語―日本語機械翻訳を試み、その精度を検討する。

2. 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)]。

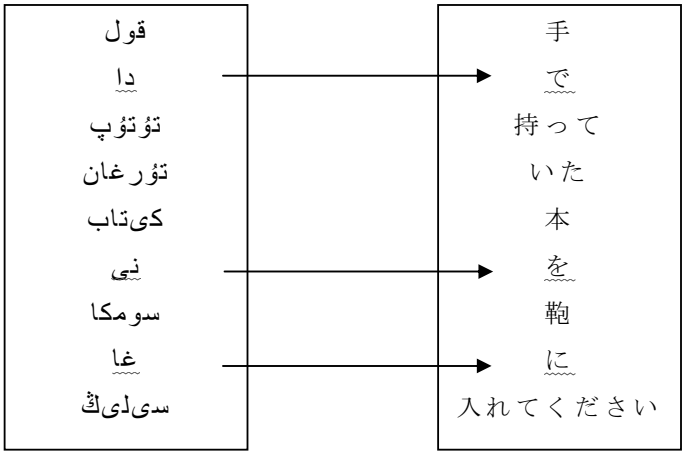
例えば:



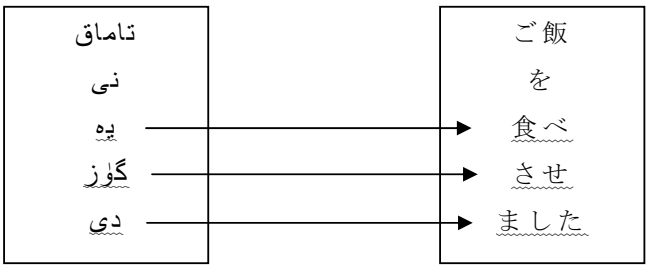
文節の文中での順序が一致する

第二に、ウイグル語と日本語は共に格助詞及び動詞接辞が存在し、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:



格助詞が存在する



動詞接辞が存在する

上記の例では、両言語の類似点を説明するために、ウイグル語の格助詞を一部分離して表記している。正書法では、先行する名詞に続けて表記する。動詞接辞も分離せずに表記する。

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。

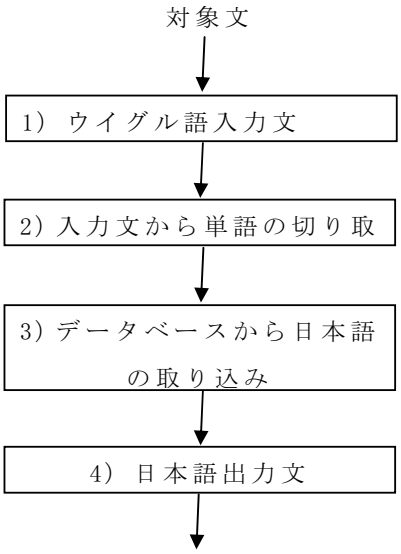


図 1 逐語訳プロセス

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。

このような相違点は、データベースの作成や訳文の質に影響すると予想される。

3. 実験用データベースの作成

現代ウイグル語で使われている単語は、語構造に基づいて大きく単一語、合成語、省略語にわけられる[2)]。今回の研究で使うデータベースは実験用データベースであるので、単一語と合成語のみを対象にした。

言語学的にウイグル語の品詞種類は名詞、動詞、形容詞、数、副詞、接続詞、感動詞、代名詞、助詞など 12 種類に分類されるが、実験用データベースは、文章の主な部分を構成する動詞、名詞、形容詞、代名詞の 4 種類の品詞の単語と記号から構成する。ウイグル語と日本語の符号ではいくつかの違いがある。例えば、日本語で使われる符号「」

をウイグル語では使わない。この符号に対してウイグル語では“”を使う。このため、色々な符号も実験用データベースに登録した。

表1に示す構造を持つデータベースを、Access2007で構成した。Windows Vistaではウイグル語を直接入力することができるが、使用できるフォントは一種類しかないため、フォントの多い無料のウイグル語入力のソフトALKATIP5.5を利用して入力した。

フィールド名	データ型	データサイズ	説明
ID	オートナンバー型	長整数型	
Uyghur（ウイグル語）	テキスト型	50文字	ウイグル語単語
Yapun（日本語）	テキスト型	20文字	日本語訳語

表1　実験用データベースの構造

実験用データベース作成に当たって直面した問題点

（1）表現が異なっているが同じ概念を表わす語が存在する場合。例えば、ウイグル語“ئامېرىكا”を表わす日本語は「米国」と「アメリカ」、“دۈشەنبە”を表わす日本語は「月曜日」と「月曜」がある。今回の実験データベースでは、どれか1つだけを選んだ。

（2）同音異義語。例えば、ウイグル語の単語“پەش”は少なくても、「上着の裾」、「骨抜き肉」、「符号の“,”」、「猫を驚かせる音」などの4つの意味を表す。今回の実験データベースでは、原文の意味に対応する日本語を選んだ。

（3）ウイグル語には日本語の普通語、丁寧語、尊敬語に対応する表現がある。例えば、ウイグル語の“دادا”、“دادىگەز”、“دادىلىرى”に対する日本語の「父」、「お父さん」、「お父様」など。これで、同じ単語を三回（普通語、丁寧語、尊敬語）入力しなければならない。そうすると、データベースは3倍大きくなってしまう。今回の実験データベースではすべて丁寧語を用いた。

（4）ウイグル語単語を派生した時の日本語翻訳の正しさ。データベースを作る時、人がウイグル語を日本語に翻訳するので、人による翻訳の正しさは機械翻訳の品質に直接影響する。

さきに述べたように、ウイグル語の単語には派生型が多いので、ウイグル語―日本語機械翻訳に使用する辞書データベースは逐語訳の手法をとる限り極めて大きくなる。この問題を避ける方法を考えることは今後の課題である。

4．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－日本語機械翻訳実行画面

5. 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.”

「明日学校へ行きます。」

5)“مەن ئەتە مەكتەپكە بارىمەن.”

「私明日学校へ行きます。」(「私は…」とした方が自然な日本語になる。)

6)“مەن ئەتە سەھەر مەكتەپكە بارىمەن.”

「私明日朝学校へ行きます。」(同上)

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،كۇتۇپخانىدا ئوقۇيمەن.”

「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。アンケート調査などによれば、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。

6．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[3)]。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。

7. まとめ

ウイグル語と日本語は言語構造が似ているため、逐語訳でも良質なウイグル語-日本語翻訳ができることが分かった。また、逐語訳に伴ういくつかの問題点を指摘した。ウイグル語の単語に派生形が非常に多いことが原因で、データベースが極めて大きくなってしまう点が特に問題で、今後の研究により解決したい。

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「ウイグル語の複合語と文節の構造について」、言語処理学会第 14 回年次大会発表論文集(NLP2008)
3)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)

謝辞

本論文は、朝日大学経営学研究科博士前記課程の論文「逐語訳によるウイグル語―日本語機械翻訳の研究」を要約加筆したものである。元論文を作成するに当たり指導をいただいた、岡本紘昭教授、板谷雄二教授、服部徳秀教授に感謝の意を表します。

Maihemutijiang Maimaitijiang （経営学研究科博士後期課程）