日本語ーロシア語機械翻訳プロジェクト JARAP

Kostyrkin Alexander, Shalyapina Zoya, Kanovich Max,
Modina Larisa, Panina Anna, Tarasova Ekaterina
ロシア科学アカデミー・東洋科学研究所

1. はじめに

JARAPは 1994 年にモスクワの東洋科学研究所が開始した実験的なプロジェクトである[1, 2]。理論面では、和露翻訳における問題点を把握し、その原因と解決法を探ることを目的としている。応用面では、理論的な研究成果を踏まえ、翻訳プログラムを作成する。言語データが翻訳に不可欠であるため、日本語コーパスの収集も本プロジェクトの目的のひとつである。この発表では、翻訳システムの基本的な処理の流れを紹介する。

JARAPはトランスファー方式のシステムであり、翻訳のプロセスは、入力文章分析・トランスファー・出力文章生成の三段階からなる。それぞれの処理は個別のプログラムモジュールによって行われる。開発中のシステムであるため、今のところ、処理しやすい統語構造の翻訳に限定している。

2. 日本語分析

入力文章の分析においては、文章を文に分け、各文を文字レベルで分解する。分解境界を特定するのに、日本語とローマ字の境界、カタカナとひらがなの境界、助詞の「を」、句点、括弧記号、引用符などを用いる。分解された要素は語彙辞書と文法辞書を用いて形態素に解析される。解析によって得られた形態素のうち一つの候補を選ぶにあたっては、「合成による分析法」を用いる。つまり、解析した形態素候補を合成してみて、文に使われている形に変化できるかを確かめ、変化できなければ、候補が誤りとされる。候補が複数であれば、前後の形態素と結合可能な候補だけを選択する。もちろん、この方法で必ずしも唯一の解釈が得られるとは限らない。入力文の形態分析の例を表1に示す。

A.文字レベル分解	画像	の	中身	は	機械 翻訳		の	対象にならない				。
B.形態素分解	画像	の	中身	は	機械	翻訳	の	対象	に	なら	な｜い	。
C.語彙素解析	гадзоо$_1$	но$_3$	наками$_1$	ва$_{2_1}$	кикаи$_1$	хон'яку$_1$	но$_3$	таисео$_2$	ни$_{2_2}$	нару$_{1_1}$ *:нс*	наи$_{1_1}$ *:зс.ин.нв*	кртчк$_1$
D.訳語	изображение *:род-п*		содер-жание	∅	машина	перевод *:род-п*		объект *:ед.тв-п*		стать	не *:наст*	.

表1. 「画像の中身は機械翻訳の対象にならない。」の文字レベル分解、形態素分解、直訳の例。曖昧な解釈は省略した。B は形態素分解の結果。C は、対応のロシア語の語彙素。番号は語義番号である。кртчк は句点を示す。コロンが先頭に立つ表記は形態素解析で求めた文法属性。нс＝未然形、зс＝終止形、ин.＝直説法、нв.＝非過去。D は、日本語の直訳であるロシア語語義とその文法属性。род-п＝生格、ед＝単数、тв-п＝造格、наст＝現在形。

得られた語彙素は、統語分析される。本プロジェクトは Z.M.シャリャピナ氏によって提唱された Entity-Based Approach（Сущностный подход）[3]に基づいている。このアプローチの基盤になっているのは、**構造的結合価**（structural valence）という概念で、他の統語論の結合価と比べると、より広く、普遍的に定義されるものである。構造的結合価は、語と語の共起関係（統語関係を含めて）を記述するためのもので、言い換えれば、ある語と他の語との意味および構造上の結合可能性を予測するものである。両語が互いに出し合う予測を満たせば統語関係が特定し得る。

構造的結合価は、動詞だけではなく、品詞を問わず、言語のあらゆる成分に適応される。つまり、構造的結合価は、単語、形態素、語句、文でも持つことができるのである。統語レベルの構造的結合価は深層格（deep case）を拡張したものと考えて良い。こうした構造的結合価を使用して、補部との関係だけでなく、付加部（状況成分と修飾成分）との関係も記述することができる。

もう一つの拡張は、結合価の持ち主が係り先であろうと、係り元であろうと、結合価の働きには変わりがないとされている。例えば、「車が走っている」の「走る」と「車」との間の関係を考えてみると、「走る」の『動作主』という結合価が『走れる動作主』を要求しており、「車」によって埋められる。これに対して「走っている車」がどう違うか考えてみれば、全く同じ語が含まれていて、全く同じ意味関係をなしていると言える。従って、両者に「走る」と「車」が同じ結合価で結ばれており、前者の結合価の持ち主が係り先で、後者の結合価の持ち主が係り元である。この例から分かるように、他の係り受け関係（「青い空」と「空が青い」、「機械翻訳」と「翻訳をする機械」等）も同じように記述することができる。

本システムの統語分析は開発中ため、現在の機能においては、隣接する「主語－述語」と「目的語－述語」という係り受け関係だけが特定できる。

【文グループ:
　【「は」グループ:
　　【画像】
　　【の】
　　【中身】
　　【は】
　】
　【述語グループ:
　　【機械翻訳】
　　【の】
　　【対象に】
　　【なら】
　　【ない】
　　【。】
　】
】

図 1. グループ分けの結果

3. トランスファー

トランスファーは、グループ分けと語順変換と語彙変換の三段階からなっている。**グループ分け**は、文を「変換単位グループ」に分割するのこと。変換単位グループ とは、隣接する成分で、翻訳変換の際にひとまとまりとして扱われ、同じような処理される成分のグループのことである。例として、名詞とそれを修飾する形容詞、動詞とそれを修飾する副詞が挙げられる。このグループは階層構造をなしている。例えば、表 1 の文の処理では、多成分の【文グループ】、【「は」グループ】、【述語グループ】が作成され、その中に入る語彙グループも作成される（図1）。

グループ分けが済んだ後、グループを構成する成分に語彙的・文法的属性を持たせ、グループの中から主要部を決める。次に、語順変換を行う。**語順変換**は、前段階で求めたグループを対象にし、そのグループとグループ内の語順を変更させるか、否かを判断する。変更するとしたグループの語順を逆にする。

ロシア語の語順は日本語とは大きく異なる。例えば、ロシア語では基本語順は SVO（主語・述語・目

的語）であり、係り元が係り先の前にも後にもあらわれることがあり、日本語の後置詞がロシア語の前置詞に当たる等の違いがある。ロシア語の語順は伝達機能と密接に関連しているため、日本語文の「は」「が」等で表現されているテーマ・レーマ構造を、ロシア語では次の例で示すように、語彙ではなく語順で表す。

（1） これは本だ　Это(これ) — книга(本)
（2） これが本だ　Книга(本) — это(これ)

規則としてロシア語の目的語は述語の後に来るため、名詞連鎖（例 3）のときも、文法関係の標識があるときも（例 4）、日本語の語順は逆になる。

（3） 言語処理　обработка(処理) языка(言語の)
（4） 言語を処理する　обрабатывает(処理する) язык(言語を)

日本語の語順のままにしておくのは、形容詞＋名詞（例 5）、副詞＋動詞（例 6）、数詞句、ラテン文字のグループである。

（5） 難しい問題　сложная(難しい) проблема(問題)
（6） ちゃんと覚える　как следует(ちゃんと) запоминает(覚える)

語彙変換では、日本語語彙をロシア語語彙に変換させる。文法的な文脈により、異なる品詞のロシア語に変換することもある。例えば、以下の三つの例の「翻訳」と言う語をロシア語の同じ意味をもつ名詞、形容詞、動詞で訳す。

（7） 完全翻訳　полный перевод(名詞)
（8） 翻訳本　переводная(形容詞) книга
（9） 翻訳する　переводить(動詞)

つまり、日本語の一義語を無理して多義語として扱うのではなく、文脈に合わせたロシア語訳を与える方法である。

4. ロシア語訳の生成

生成モジュールの入力は、各訳語の辞書形とその文法範疇からなっている。生成過程は入力の修正から始まる。修正されるのは以下のものである。

語彙とその語彙に与えられた文法属性が一致するか確認する。たとえば、トランスファーにおいて、名詞が「時制」属性を、又は動詞が「過去」と「現在」属性ををを与えられてしまうなどように、一致がみられないところでは、文脈によって品詞を変更させるか、一つの語彙を語結合に置き換えるか、矛盾する属性のうちからどちらかを削除するかという対策をとる。例えば「бюрократ（官僚）」を動詞化させようという場合、ロシア語に派生語がないため、語結合の「действовать как бюрократ（官僚のような行為をする）」に置き換える。

修正が済んだ後、形態生成を行う。生成の際は以下の様な補助的な処理をおこなう。

係り先が係り元の「性」「数」「格」を求める場合、それを係り元に付加する。ロシア語の構造が要求する前置詞と接続詞を挿入する。主語と目的語の有無によって、述語が能動態なのか受動態なのか選ぶ。トランスファーから必須の文法情報が得られなかったとき、デフォルト設定に基づいて補う。たとえば、ロシア語の動詞の活用には「人称」と「数」を表す必要があるが、日本語の入力文からそれが不明な場合、三人称単数にする。与えられた語彙と要求する語彙が補充法の関係にある場合は、語彙そのものを変更する。例えば、「人たち」であれば、内部表現が「человек:мн（人：複数）」であって、複数形を「люди」に変更させて生成する。

содержание:NN{1_2}.им-п.R
G1[изображение:NN{1_4}.род-п]
изображение:NN{1_4}.род-п.R
D1[содержание:NN{1_2}.им-п]
не:NN{1_6}.наст
стать:NN{1_8}.R
G1[содержание:NN{1_2}.им-п]
объект:NN{1_9}.ед.тв-п
машинный:NN{1_12}
перевод:NN{1_13}.род-п
.:NN{1_14}.|

図 2. 形態生成の入力の例. NN は語形の訳文の中での位置. G(=governor)は係り先を, D(=dependent)は係り元を、番号は結合価の番号を、「｜」は文末を意味する. 他の表記は表 1 と一緒.

生成の結果は入力文に相当するロシア語の文を出力する。
表1に挙げた文を JARAP で訳した結果は以下のようである。

（10） Содержание изображения не становится объектом машинного перевода.
（中身 画像の）が ない なる 対象に （機械 翻訳）の

例文に「は」を入れて、「画像の中身は機械翻訳の対象にはならない。」の訳を求めると次のようになる。

（11） В объекте машинного перевода содержание изображения не состоит.
対象に （機械 翻訳）の （中身 画像の）が ない なる

参考文献

[1] Modina L.S. & Shalyapina Z.M. – The JaRAP experimental system of Japanese-Russian automatic translation // Proceedings of COLING 94. I. Kyoto, p.112-114, 1994.

[2] Шаляпина З.М. Система японско-русского автоматического перевода ЯРАП/1: первые экспериментальные результаты (和露機械翻訳システム JARAP/1 の初実験の成果). Бюллетень (Newsletter) Общества востоковедов РАН. 10. Москва: ИВ РАН, с.164-226, 2004.

[3] Shalyapina Z.M. Computational approaches: An entity-based linguistic framework // Paris Lectures in Japanese Linguistics, Ed. André Wlodarczyk, p.167-211, Kuroshio, 2005.