機械翻訳エンジン jaw における日本語パターンの記述形式及びパターン照合処理について

浅井 良信　穆 貴彬　玉置 健二　池田 尚志
岐阜大学 工学部

1. はじめに

我々は、日本語からいろいろの言語への翻訳を行う機械翻訳エンジン jaw を開発している。現在、中国語、シンハラ語、ベトナム語、ミャンマー語日本手話を目的言語とする翻訳システムの構築を試みている[2][3]。

jaw による翻訳実験を進めながらシステムの開発を進めてきた。日本語パターンについても、多段階のパターンや機能語部分に条件を持たせたパターンも記述できるようになり、結合価レベルのパターンのみではなく大抵の大域的なパターンも記述できるようになった。本報告では、現在 jaw で用いている日本語パターンの記述形式とパターン照合処理のアルゴリズムを中心に述べる。

2. 機械翻訳エンジン jaw の概要

機械翻訳エンジン jaw は VC++で作成されたトランスファー方式の機械翻訳エンジンである。日本語のパターンを使って、日本語の構文構造を翻訳対象の言語の表現構造に置き換えて翻訳を行なう。

jaw の基本的な処理の流れを図 1、jaw のインターフェースを図 2 に示す。

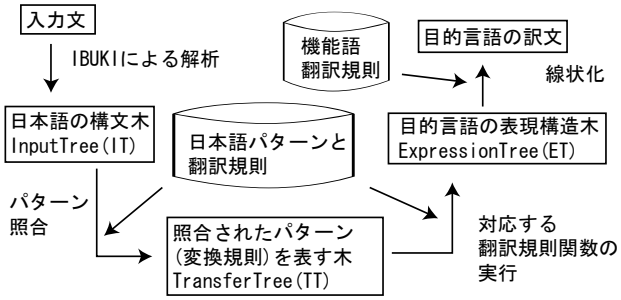

図 1　機械翻訳エンジン jaw の処理の流れ

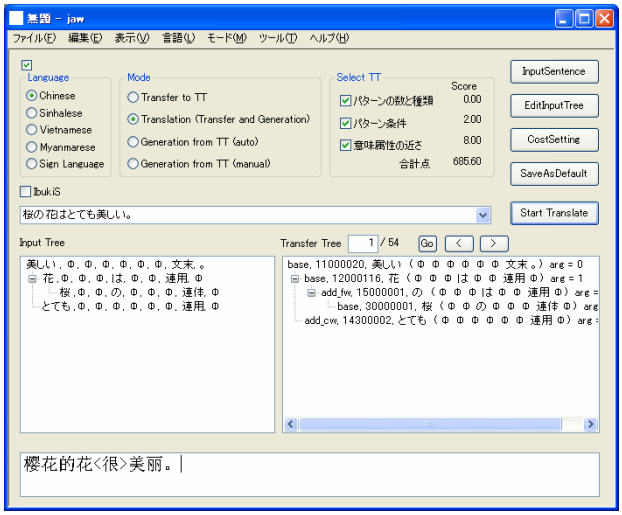

図 2　jaw のインターフェース

3. 日本語パターンの記述形式

3.1. 日本語解析と日本語パターン情報

jaw ではまず入力文を、文節を節とし、係り先文節を親とする木構造(IT : InputTree)で表現する。日本語解析は我々の研究室で開発している日本語解析システム IBUKI[1]を使って行なう。IT の節はこれらの解析情報を持つ。

```
桜の　{文節:1／係り先:4／品詞カテゴリー:N／
       CW:桜／FW1:Φ／FW2:Φ／FW3:の／
       FW4:Φ／FW5:Φ／FW6:Φ／係り:連体／句:Φ}
花は　{文節:2／係り先:4／品詞カテゴリー:N／
       CW:花／FW1:Φ／FW2:Φ／FW3:Φ／
       FW4:は／FW5:Φ／FW6:Φ／係り:連用／句:Φ}
とても {文節:3／係り先:4／品詞カテゴリー:AV／
       CW:とても／FW1:Φ／FW2:Φ／FW3:Φ／
       FW4:Φ／FW5:Φ／FW6:Φ／係り:連用／句:Φ}
美しい。{文節:4／係り先:0／品詞カテゴリー:P3／
       CW:美しい／FW1:Φ／FW2:Φ／FW3:Φ／
       FW4:Φ／FW5:Φ／FW6:Φ／係り:文末／句:。}
```

図 3　IBUKI の解析：「桜の花はとても美しい。」

jaw は日本語パターンと IT を照合し命題的内容の翻訳を行なう。日本語パターンとは日本語の表現から文の構造や単語を取り出し、一般形として表したものである。日本語パターンはリレーショナルデータベース(RDB)上に記述されており、図 4 のような情報を持っている。

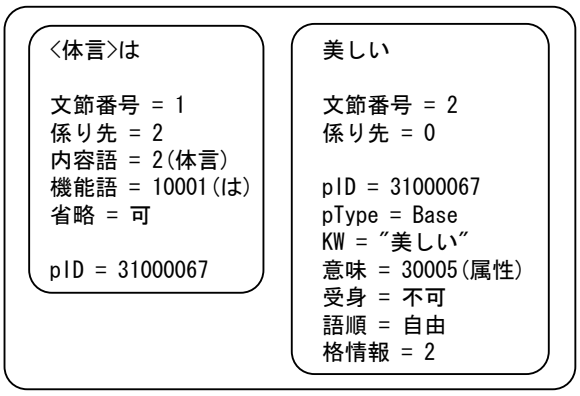

図 4　日本語パターン例：<体言>は美しい

3.2. 日本語パターンの型

jaw では Base、AdditionCW、AdditionFW の 3 種類の日本語パターンを設けている。日本語パターン及び対応する翻訳規則を作成するためのエディターjawEditor の開発も行なっている。

Base 型日本語パターン

内容語を検索キーワード(KW)とし、単語や節を表現するパターン。「私」、「彼」、「家」などの名詞や、「<主体>が美しい」、「<人>が<場所>に泊まる」などの述語節を Base 型の日本語パターンで記述する。

係り受けが 2 段階以上の Base 型日本語パターンも作成することができる。「<体言>が<人>の口に合う」や「<人>が<用言>と言って誉める」などは 2 段階の係り受けを持っているが、この様な固有表現も日本語パターンで表すことができる。

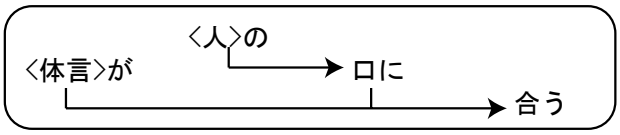

図 5　2 段階の係り受けを持つ日本語パターン

AdditionCW 型日本語パターン

Addition 型の日本語パターンは Base 型日本語パターンの文節に係る文節を表現する。

AddtionCW 型の日本語パターンは、内容語を KW とする接続文節を表すパターンである。「大きい<体言>」、「とても<用言>」などの形容詞・副詞や「<体言>のあおりを食って<用言>」などの文節にかかる固定表現を AddtionCW 型の日本語パターンで記述する。

Addition 型日本語パターンでも 2 段階以上の係り受けを持つものを作成できる。「<体言>のあおりを食って<用言>」は 2 段階の係り受けを持つ Addition 型の日本語パターンの例である。

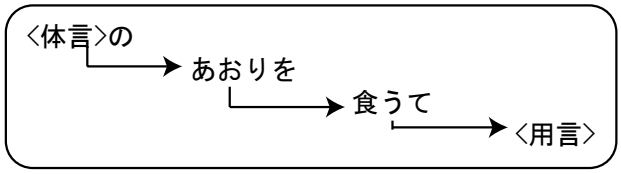

図 6　2 段階の AdditionCW 型日本語パターン

AdditionFW 型日本語パターン

機能語を KW とする接続節を表すパターン。「<体言>の<体言>」、「<用言>て<用言>」などの助詞による文節接続や「何を<用言>ても<用言>」などの従属節を表すようなパターンを AddtionFW 型の日本語パターンで記述する。

jaw は日本語パターンの情報を木構造(PT : PatternTree)で表現し、入力文の木構造(IT)との間でパターン照合を行なう。

日本語パターンには、翻訳規則が対応づけられている。翻訳規則は、目的言語の対応する表現構造(ET : ExpressionTree)を作り上げるプログラムである。ET は C++のオブジェクトネットワークであり、その部品や要素は目的言語毎に設計している。

jaw では日本語パターンと翻訳規則を使って命題的内容の翻訳が行なわれる。

4. パターン照合のアルゴリズム

4.1. パターン照合処理

照合処理はITの根の節(通常文では文の述語文節)から行い、Base 型パターンの照合、Addition 型パターンの照合の順に行なう。

照合結果は、木構造(TT：TransferTree)で表される。これはどの日本語パターンで照合されたかを表現するもので、それぞれの節が入力文解析の情報や照合された日本語パターン及び対応付けられた翻訳規則などの情報を持っている。TT 上の翻訳規則を実行して、目的言語の表現構造 ET を作成する。

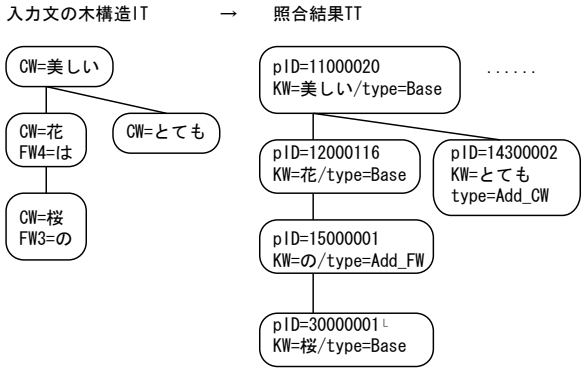

図 4　IT と TT 例：「桜の花はとても美しい」

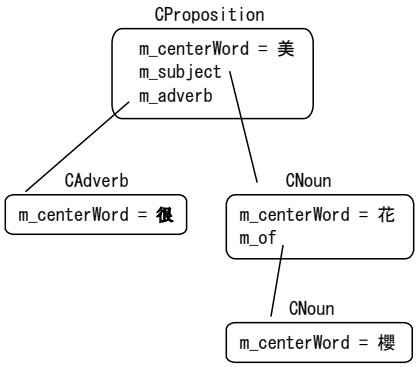

図 5　中国語の ET 例

4.2. Base 型パターン照合

照合対象の IT の内容語を KW として Base 型日本語パターンの検索を行なう。検索されたすべての日本語パターンに対して照合を行なう。

Base 型パターン照合時に、字面情報(フラグ)が与えられていたとき、機能語・内容語条件の無い単文節の日本語パターンが 1 つ検索されたこととして以下の Base 型パターン照合を行なう。

条件文節情報が与えられていたとき、検索された日本語パターンに条件文節を連結した日本語パターンを使って Base 型パターン照合を行なう。

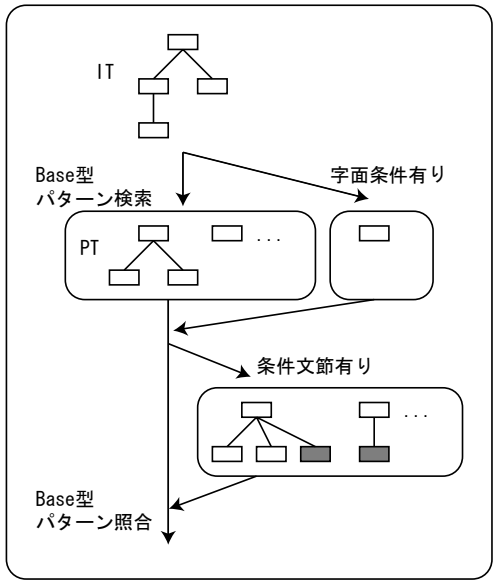

図 5　Base 型パターン検索

個々のパターン照合は、名詞(「私」、「花」など)のような 1 文節の日本語パターンの場合、その日本語パターンの情報を持った TT を結果とする。

複数文節(「<体言>は美しい」など)の日本語パターンの場合、次の手順で処理を行なう。以下の処理で、係り受け 1 段階ずつ条件照合していく。

1. 入力文 IT のそれぞれの係る文節が、日本語パターンの各文節の機能語条件と内容語の字面条件を満たすかどうか調べる。
   日本語パターンが照合対象に当てはまらないとき、このパターンの照合を終了する。
2. 機能語条件を満たす文節の IT を対象とした照合処理を行なう。照合結果のうち内容語条件を満たすものを記録する。
   係る文節の IT 照合処理で日本語パターンの対応する文節の内容語条件が字面で与えられていた場合、係り文節の照合処理に字面情報を与え照合処理を行なう。
   対応する日本語パターンが 2 階層以上の係り受けを持つ場合(「<人>が<用言>と言って誉める」など)、2 階層以上の部分(例では「<用言>と」)を条件文節情報として与え照合処理を行なう。

3. 日本語パターンの各文節に対応する TT が得られたら、係る文節の TT のすべての組み合わせを取りパターンの KW 文節の情報を持った TT に接続し結果とする。

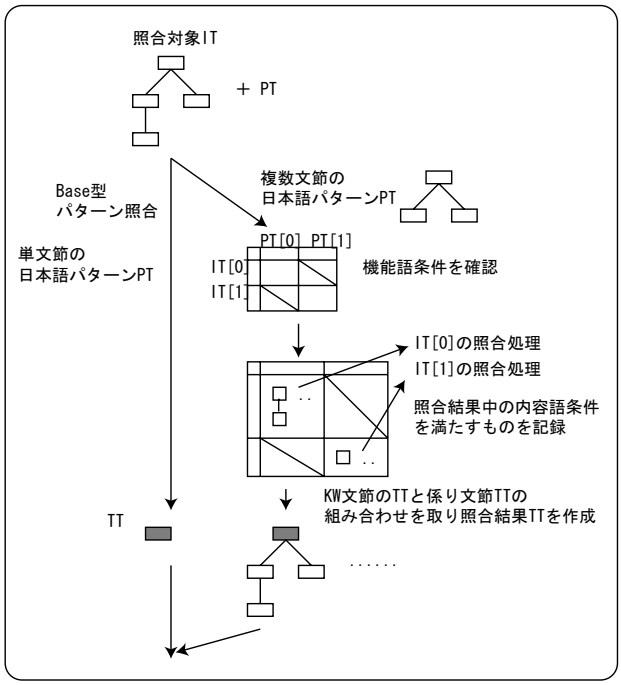

図 6　Base 型パターン照合

4.3. Addition 型パターン照合

Addition 型パターン照合は、Base 型パターン照合の未照合文節に対して行なわれる。IT の未照合文節を 1 文節ずつ取り出して Addition 型照合を行なっていく。未照合文節が全て照合できたら、照合成功とし得られた TT を照合結果とする。

Addition 型パターン照合には、AdditionCW 型パターン照合、AdditionFW 型パターン照合、内の関係の用言連体修飾の照合、接続照合(外の関係の用言連体修飾を照合できる)があり、この順で照合が行なわれる。このうち、AdditionCW 型パターン照合、AdditionFW 型パターン照合について述べる。

AdditionCW 型パターン照合

照合対象文節の内容語を KW として Addition CW 型パターンの検索を行ない、検索された全てのパターンに対して照合を行なう。照合では、まず、係り先文節の内容語条件･機能語条件を満たすかどうかを確認する。条件を満たしていたら、Base 型パターン照合と同じ手順で KW 文節とその係り文節の照合を行なう。得られた TT を係り先 TT と接続詞結果とする。

AdditionFW 型パターン照合

照合対象文節の機能語を KW として Addition FW 型パターン検索を行なう。AdditionCW 型パターン照合と同様に、係り先の条件を満たすか確認する。この後、照合対象文節の照合処理を行なう。照合結果の TT を受け取り、内容語条件を満たすものを選択、KW 文節の情報を持った TT に接続し、それを係り先 TT と接続し結果とする。

「何を<用言>ても<用言>」などの従属節を表すような日本語パターンの場合は、2 階層以上の部分(例では「何を」の部分)を条件文節情報として与え照合処理を行なう。

5. 翻訳規則(表現構造)と線状化

照合で作成された TT を使って翻訳規則を適用し、目的言語の表現構造木(ET)を作成する。jaw では表現構造が作成された後に機能語翻訳を行ってから、線状化処理(要素並び替えなどを行なう処理)を行ない、訳語を出力する[2]。

6. おわりに

現在の jaw の日本語パターンの記述形式とパターン照合処理について述べた。大抵の大域パターンについては扱えるようになっているが、未だ残された問題もある。今後はそれらの問題の解決を行なうとともに、例文の翻訳を行ないながら新しい問題を探し対処していくつもりである。

参考文献

[1] 日本語文解析システム IBUKIC/S について : 山田佳裕他：言語処理学会第 12 回年次大会(2006)
[2] 日本語から他言語への機械翻訳エンジン jaw : 宇野修一他：言語処理学会第 11 回年次大会(2005)
[3] 機械翻訳システム jaw と他言語への翻訳実験 : 浅井良信他：言語処理学会第 11 回年次大会(2005)