

自然言語処理システムのための 日本語文生成ツールの試作

4T-7

重永 信一 安川 秀樹

松下電器産業(株) 技術本部 システム東京研究室

0. はじめに

ICOTにおいて日本語処理の基礎技術である日本語解析や日本語文生成を共通ツールとして整備しようという試み(LTB)が進められている。本試作はこのLTBの一環として行ったものである。本システムは、文法的な情報のみを用いて、標準的な日本語文を生成する。システムの入力形式として、文法的情報の集合を表現する生成用中間表現形式を設定する。試作では、予め用意した例文に対して中間表現データを与えて生成実験を行なうとともに機械翻訳システムの文生成部に組み込んで、生成機能と設定した中間表現形式の妥当性を評価した。

1. システム構成

図1にシステム構成と処理の流れを示す。生成エンジンは、生成用辞書を参照しながら、与えられた中間表現データに生成用文法を適用して語彙の確定と語順の決定を行う。生成エンジンの処理の結果は、形態素生成部に渡され、形態素生成処理を経て日本語表層文として出力される。辞書と文法は、トランスレータにより生成エンジンが実行可能なCILの述語に変換される。

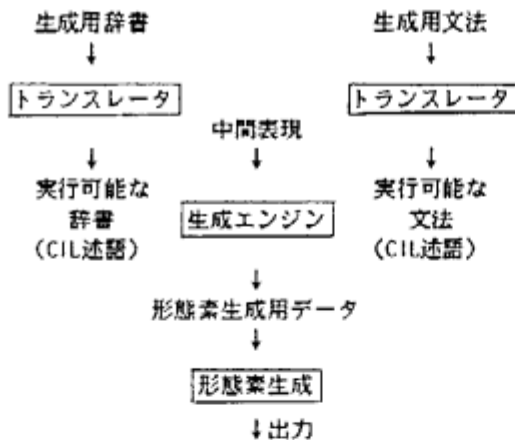


図1 システム構成

2. 日本語生成用中間表現

2.1 日本語生成用中間表現の特徴

日本語生成用中間表現は以下の特徴を持つ。

- (1) 日本語文の構文的情報を属性-属性値対からなる属性構造(feature structure)として表わす。
- (2) 表層表現を規定するために必要な属性が全て与えら

れている。

- (3) 語順は規定されていない。
- (4) 代名詞化、提題化、省略、態の指示が明記されている。

2.2 属性構造

属性構造は属性-属性値対の集合として定義される。

ここでは、日本語文の構文的情報を表すために、以下の体系を設定している。

```

<文> ::= [ <接続詞> ] <文核>
<文核> ::= ' <単文> | <重文>
<単文> ::= <語句成分> [ <主題属性> ]
<重文> ::= <等位接続>
           <単文>
           <文核>
<語句成分> ::= <語義属性>
               [ <格要素>... ]
               [ <文法的属性>... ]
               [ <接続関係> ]
<格要素> ::= <深層格名> <語句成分>
  
```

以下に、システムで用意している属性の種類を示す。

- (1) 文法的機能
 - ・格要素
 - ・接続関係
 - 連用修飾、連体修飾、従属接続
 - ・接続詞
 - ・主題
 - 格支配のない主題化に対処するための入れ物。
- (2) 語義属性
 - 生成用辞書を引くための属性。
- (3) 語義に付随した属性
 - a) 名詞類が持つもの
 - 数、人称、性、意味カテゴリ、代名詞化、限定詞(この、その、あの、どの)
 - 提題化、格支配、省略
 - b) 述語が持つもの
 - 名詞化、時制、アスペクト、ムード、態度、否定、受動化
- (4) 等位接続
- (5) 接続形態
 - 連体修飾や従属接続の形態を与える。

「自然の中には様々な物質が存在している。」という文を生成するための中間表現の記述例を図2に示す。

{語義/存在する,
アスペクト/状態,
対象/{語義/物質,
連体/{格支配/対象,
語義/様々,
対象/{語義/物質}}},
場所/{語義/中,
連体/{形態/の(),
語義/自然},
提題化/yes}}.

図2 中間表現データの記述例

3. 生成用辞書

生成用辞書には、語彙の確定、形態素生成、格要素の語順の決定などを行うのに必要な単語ごとの情報が記述される。辞書の記述形式は次の通りである。属性列には、個々の単語の文法的な属性が属性-属性値の組として記述される。

品詞名(<エントリ名>)
'=>' { <属性列> } .

4. 生成用文法

生成用文法は、以下のフォーマットに従って、句構造文法形式に記述される。

<文法規則> ::= <規則名> (<引数1> , <引数2>)
'with' [<条件部> ,]
<結果部>
['=>' <下位規則列>] .
<引数1> ::= 中間表現の部分構造に対応する変数
または値
<引数2> ::= 結果部に対応する変数
<条件部> ::= CIL述語の列
<結果部> ::= 'lex' ':' <prologの項> |
'pattern' ':' (<変数列>)

条件部には、下位規則に与えられるべき構造変数の属性値の並び、下位規則の適用に必要な処理手続き、適用条件をチェックするための手続きなどが記述できる。結果部における ':' は規則の適用結果として、規則の第2変数にlex値あるいはpattern値として右辺部がユニファイされることを示す。

5. 生成エンジン

生成エンジンは、入力される中間表現データに辞書及び文法を適用して、表層文の語彙と語順を確定する。文法の適用結果は表層文の語彙情報の列であり、文法規則の適用が完了した後で、規則の第2変数にバインドされたpattern値あるいはlex値を取り出すことによって得られる。生成用の辞書は、語彙の確定、述語の格配列の決定、文法規則を適用する際の個々の単語の文法的なチェックのために参照される。生成エンジンで行われる主な処理は次の通りである。

- (1) 中間表現データに対する文法の適用
- (2) 結果部の取り出し
- (3) 形態素生成のためのデータの構成
- (4) 出力処理

(5) 文法規則中の手続きの処理

- ・文法規則適用のためのチェック
- ・格成分の取り出し
- ・格要素、修飾句などの語順の決定
- ・代名詞化、名詞化、アスペクトの処理、モダリティの処理

6. 形態素生成

形態素生成部では、生成エンジンの出力結果に対して、品詞や接続情報を使って次のような処理を行う。

- (1) 活用処理
- (2) 連続した助詞の処理
- (3) 助動詞の語彙変形
- (4) 音便の処理

7. 今後の課題

今後の検討課題として次のようなものがある。

- (1) 共起性の扱い
語彙の間の共起性の扱いと記述の枠組みの検討。
- (2) モダリティの扱い
複合したモダリティを扱うためのモダリティのスコープの扱いの検討。
- (3) アスペクトの扱い
複合アスペクトの扱いの検討が必要。
- (4) 丁寧表現の処理
丁寧表現に伴う語彙変化の扱いの検討。
- (5) 補文標識の使い分け
個々の動詞により定まる補文標識の使い分け、構文的に規定される補文標識の扱いの検討。
- (6) 受動態
[木村 85] の助詞変換の型の分類などを参考にし
たより正確な助詞変換の実現。

今後は、上記の課題をクリアしながら、汎用ツールとしての機能を充実させるために、文法や辞書のトランスレータの機能強化、文法のデバッグ機能といった開発環境の整備を含めた総合的なツール作りを行う予定である。

8. まとめ

選定した例文に対する中間表現データの記述にあたっては、滑らかな日本文となるように意識的に語義及びその他の属性の値を与えた。本来は、前段階のシステムから何らかの意味表現が与えられ、その意味表現から生成するという流れを想定しているもので、試作のように都合よく中間表現が与えられる訳ではないが、ツール作りの第一歩としては、汎用ツールとして基本的な機能の妥当性を十分確認できた。

なお、本研究は、第5世代コンピュータプロジェクトの一環として、(財)新世代コンピュータ技術開発機構からの委託(発注仕様番号1801)により行った。

参考文献

[木村 85] 木村睦子, 空閑茂起, "態による格助詞変換"

計量国語学第15巻2号, 1985.