構造化された日英文型パターンを用いた線形・非線形融合型日英構造変換方式

川辺 諭† 武本 裕‡ 宮崎 正弘‡

†科学技術振興機構 ‡新潟大学大学院自然科学研究科

1 はじめに

機械翻訳においては、原言語と目的言語間で単語同士が単純に置き換えられない、非線形に対応する部分があるために、品質の良いこなれた訳文が得られないことが多い。この問題を解決するために、非線形対応が見受けられる日英対訳データに関して、語・句・節レベルで汎化作業を行い、汎用性の高い構造化された日英文型パターンを作成する。本方式では日本語英語間の翻訳処理において、入力文の構造に適合するパターンデータを検索し、得られたパターンから目的言語表現の大域的な構造を得る。パターンとの差分となる自由格成分や節などの線形要素は、適宜変換して大域的な構造へ埋め込むことで、こなれた目的言語表現を出力することが可能になる。本稿では日英文型パターンの照合方式を提案し、線形・非線形融合型の日英翻訳方式における有効性について論じる。

2 日英翻訳システムの概要

現在試作中の日英翻訳システム (図 1) の概要を述べる。

2.1 日本語文解析部

日本語形態素解析処理：日本語形態素解析システム **Maja** により、CYK 法を用いて日本語入力文を分割する。

日本語形態素解析後処理：CYK 法では処理できない 3 単語以上の形態素の連鎖に関して修正処理を施し、助詞相当語・助動詞相当語の抽出、同形語の判別などを行う。

日本語構文解析前処理：重文、複文に関して構造解析時の多義を抑制するために、節の区切りと推測される単語を越えて部分木を生成しないように、抑制情報を付加する。

日本語構文解析：日本語構文解析システム **SGLR - plus/J** により、GLR 法を用いて日本語文を構造化し、入力された日本語文の統語的な構造である “表現構造” と、意味的な構造である “認識構造” を抽出する。

2.2 変換部

木構造変換：**SGLR - plus/J** の文法は、記述量を削減して見通しを良くし、パーザの負荷を軽減するといった理由から、記述に冗長性があり、出力される木構造には、兄弟がいない冗長なノードが含まれる。変換処理に先だってこれらの不要なノードを削除し、“日本語縮退木” を得る。

日英非線型変換部：日本語の重文・複文に関して、格パターンを用いて翻訳された節同士を単純につないだだけでは英語文を生成することができない、日本語・英語表現の対を構造化した日英文型パターン対を準備して、日本語入力文と適合するパターン対を利用して英語出力文の大域的な統語構造を決定する。

日英線形変換部：日英間の表現が対応する語・句・節に関して、対照辞書や格パターンを用いて翻訳処理を行う。

日英規則用例融合変換部：日本語入力文で構造が複雑な名詞句や複合名詞に関して、規則・用例融合型の変換処理を行い、対応する英語名詞句などを生成する。

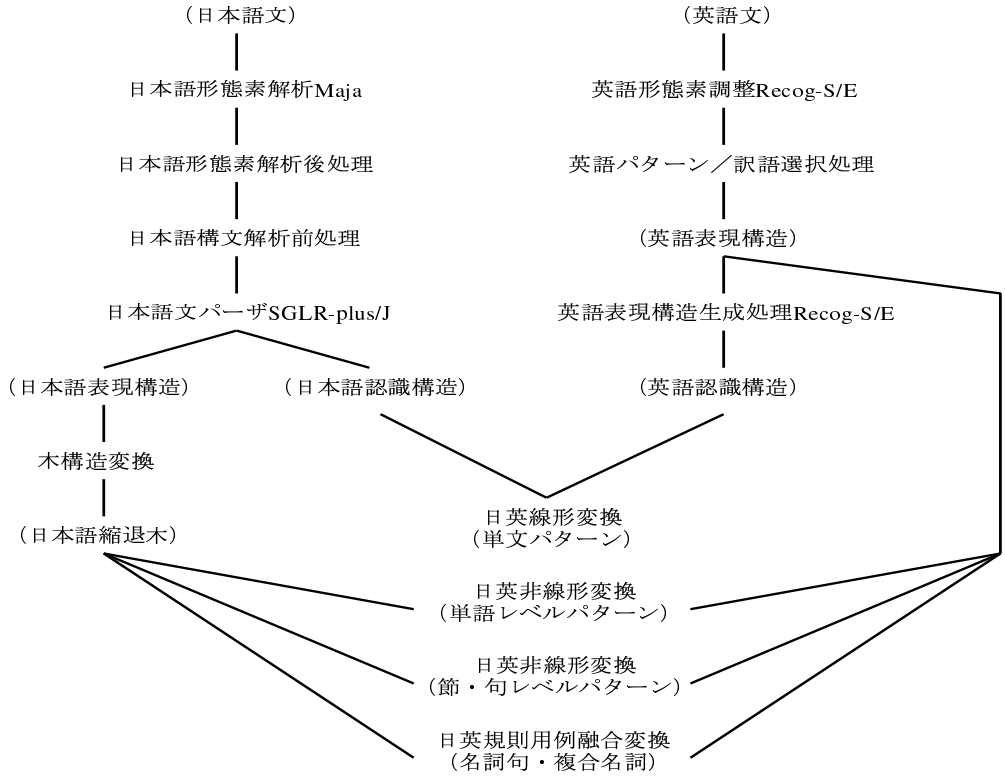

図 1: 試作中の日英翻訳システムの概要

2.3 英語文生成部

英語表現構造生成処理 Recog-S/E：日英非線型処理によって日英文型パターンから得られた英語文の大域的な構造に、日英線形処理で得られた英語単文を埋め込み、英語出力文の統語的な構造である “英語表現構造” を生成する。

英語パターン／訳語選択処理：日本語入力文の意味的な情報を利用し、訳語選択処理を行う。

英語形態素調整 Recog-S/E：英語文の名詞に冠詞を付与し、名詞・動詞・形容詞に関して語尾処理を行うなどの形態素調整を施す。

3 言語データ

日英翻訳処理システムの変換部で用いる言語データについて述べる。

3.1 単文パターン

線形変換部において節を日英翻訳するために、“単文パターン” を用いる。単文パターンは自由格成分の抽出・挿入処理のために構造化されており、また、それぞれの格要素には意味カテゴリが付加されている [1]。図 2 は「食べる」 “eat” の単文パターンである。

日本語	N1 が N2 を 食べる
英語	N1 eat N2
カテゴリ	N1：4 人 535 動物 N2：838 食料 534 生物

```
# 日本語側構造
|-s
  |-cl
    |-vp
      |-np
      | |-n-< N1 >
      | |-p-< が >
      |-vp
        |-np
        | |-n-< N2 >
        | |-p-< を >
        |-v-< 食べる >
```

```
# 英語側構造
|-s
  |-cl
    |-sbj
    | |-np-< N1 >
    |-vp
      |-v-< eat >
      |-obj
        |-np-< N2 >
```

図 2: 単文パターンの例

3.2 日英文型パターン

要素合成法では日本語表現から英語表現を生成することが難しい日英表現の対に関して、対応する語・句・節を汎化作業によって変数化し、日英表現の全体を構造化した “日英文型パターン” を準備する。図 3 は日英文型パターンの例である。

日本語	CL1 て N2 は 心が 痛む
英語	it ail N2.obj that CL1

```
# 日本語側構造
|-ss
  |-s
  | |-cl-< CL1 >
  | |-aux-< て >
  |-s
    |-cl
      |-vp
        |-np
        | |-n-< N2 >
        | |-p-< は >
        |-vp
          |-np
          | |-n-< 心 >
          | |-p-< が >
          |-v-< 痛む >

# 英語側構造
|-s
  |-cl
    |-np
    | |-pron-< it >
    |-vp
      |-v-< ail >
      |-obj
      | |-np
      |   |-n-< N2.obj >
      |-that-cl
        |-conj-< that >
        |-cl-< CL1 >
```

図 3: 日英文型パターンの例

日英文型パターン中の語・句・節変数や字面には、“ V2. 連用 ”、“ N2.obj ” などの統語情報や、“ た. 既定 ”、“ て. 原因 ” などの意味情報が付加されており、これらは日英文型パターン検索時の制約条件や、英語文生成時における形態素調整、時制、相、様相などの主体表現付与処理のための情報として利用される。

4 パターン照合処理

パターン照合部では、日英文型パターンの候補を検索し、各候補の日本語パターンと入力日本語の構造を比較することで、適用可能なパターンを選び出す。

4.1 日英文型パターン検索処理

日英文型パターン検索処理部では、日本語入力文に関して、用言・体言・接続表現などの字面・意味属性をキーとして、類似の字面・意味属性を持つ日英文型パターンを検索する。

4.2 木構造比較処理

木構造比較処理部では、日英文型パターンの日本語側構文木 (以下 PJ 木) と日本語入力文の構文木 (以下 J 木) を比較することで、日英文型パターンが適用可能かどうかを判断する。また、比較作業と並行して PJ 木と J 木の差分を抽出する。

木構造は CL や VP などの文法カテゴリに関する生成規則が左・右再帰で記述されているため (表 1)、左右の再帰的な構造になっている。

表 1: 走査関数一覧

走査関数	再帰	再帰型生成規則の例
CL 走査関数	左再帰	`cl -> cl aux`
VP 走査関数	右再帰	`vp -> [pp advp] vp`
NP 走査関数	右再帰	`np -> [ajp cl] np`
AJP 走査関数	右再帰	`ajp -> adv ajp`
AJV 走査関数	右再帰	`ajvp -> adv ajvp`
ADVP 走査関数	右再帰	`advp -> adv advp`

図 4 は PJ 木、J 木の比較処理の例である。比較処理の流れは以下の通り。

0 比較作業は根から葉に向かって行う。
1 左再帰の木構造を下方に向かって別種のノードが現れるまで走査し、木構造の右隅要素を集める。

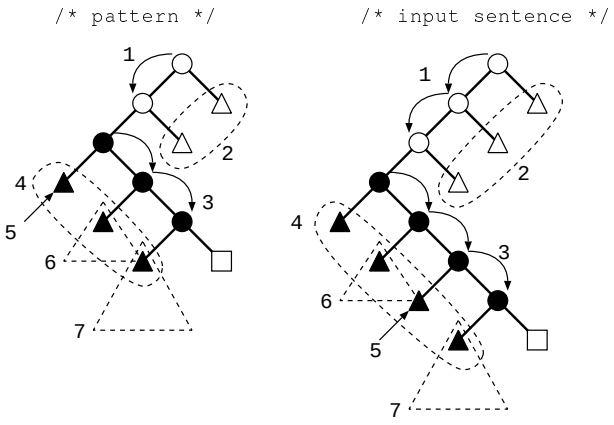

図 4: PJ 木と J 木の比較

2 1. の処理で集められた PJ 木の右隅集合 R_{PJ} と J 木の右隅集合 R_J を比較し、R_{PJ}　R_J でなければ失敗。

3 右再帰の木構造に関しても同様に、下方に向かって別種のノードが現れるまで走査し、木構造の左隅要素を集める。

4 3. の処理で集められた PJ 木の左隅集合 L_{PJ} と J 木の左隅集合 L_J を比較し、L_{PJ}　L_J でなければ失敗。

5 以上の比較処理を木構造の末端にたどり着くまで繰り返し、J 木、PJ 木の走査ポインタが同時に末端にたどり着けば成功。

右隅集合、左隅集合の比較処理によって、PJ 木、J 木の 2 つの木構造の差分となる助動詞、自由格成分などの要素が抽出される。

図 5 は、比較照合処理によって得られた大域パターンに、単文パターンを用いて変換された差分 CL を埋め込む例である。

```
■ 入力文 : 彼が試験に失敗して私は心が痛む
■ CL変換
JCL1 = 彼が 試験に 失敗する
|-cl
  |-sbj
  | |-np
  |   |-pron-< he >
  |-vp
    |-v-< failed >
    |-pp
      |-p-< in >
      |-np
        |-det-< the >
        |-n-< exam >
```

```
■ 大域構造へCLを埋め込んだ結果
|-s
  |-cl
    |-np
    | |-pron-< it >
    |-vp
      |-v-< ail >
      |-obj
      | |-np
      |   |-n-< I.obj >
      |-that-cl
        |-conj-< that >
        |-cl
          |-sbj
          | |-np
          |   |-pron-< he >
          |-vp
            |-v-< failed >
            |-pp
              |-p-< in >
              |-np
                |-det-< the >
                |-n-< exam >
```

図 5: 大域構造への埋め込み

5　おわりに

非線型な対応を持つ複文・重文表現の日英対訳データに関して、語・句・節レベルでの汎化作業と構造化作業を行った日英文型パターンを構築し、日英翻訳においてこれらのパターンを適用することで出力文の大域的な構造を決定し、よりこなれた英語文を生成する手法を提案した。日英文型パターンを拡充し、試作システムの実装を進めることが今後の課題である。

6　謝辞

この研究は、科学技術振興機構 (JST) の戦略的基礎研究事業 (CREST) の支援と、科学研究費補助金基盤研究 (B)(課題番号 13480091) を受けています。

参考文献

[1] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 : 日本語語彙大系, 岩波書店 (1997)

[2] 池原, 佐良木, 宮崎, 池田, 新田, 白井, 柴田 : 等価的類推思考の原理による機械翻訳方式, 電子情報通信学会, 信学技報 TL2002-34 (2002-12).