パターン対を用いた 線形・非線形融合型日英構造変換方式

武本 裕 * 川辺 諭 † 宮崎 正弘 *

* 新潟大学大学院自然科学研究科 † 科学技術振興事業団

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対応せず単純に置き換えられないような、対応関係が非線形の部分があるために、品質の良いこなれた訳文が得られないことが多い。

本手法では非線形対応が見られる日英対訳データをもとに作成した単文および重文・複文レベルの日英文型パターンを利用して、日英間で構造変換を行なう。線形の対応関係が成立する部分については単文レベルの日英パターン対から得られた単文を組み合わせて要素合成的に変換を行なう。これにより、非線形な対応関係のために適切な翻訳結果が得られない場合にも品質の良い訳文を得ることを狙う。

本稿では上記手法による日英機械翻訳システムの変換部を提案し、その有効性について論じる。

2 要素合成の問題点

従来の翻訳システムで行なわれている要素合成による方式では、柔軟な翻訳を行なうことはできず、不自然な翻訳結果となってしまう場合がある。重文や複文の場合、単に単文に分解してそれらを変換した後合成するということが行なわれる。例えば、「あまりに暑かったので、コートを脱いだ。」のような文に対して「so ~ that」構文を使って表現することができない。

3 本方式の基本的な考え方

ここでは、本方式の基本的な考え方について説明する。まず、原文が非線形な対応をする可能性があるかどうかを判定する。

データベースに (主に重文・複文に対応した) 非線形パターンが登録されている。

データベースに登録されている場合には、そのパターンに基づいて変換を行なう。パターンにおいて、句・節が変数化されている場合には名詞句／動詞句等の変換部、節変換部で処理する。節変換部は、単文変換パターンを利用して変換する。データベースに登録されていない場合には、従来通り単文への分解を行なって各部分を変換後、合成する。

4 認識構造

時枝誠記 [1] が提唱し、三浦つとむ [2] が発展的に継承した言語過程説によれば、言語表現を行なう表現主体である人間は、対象の客観的姿を概念として捉え (客体認識)、それに対してある種の主体的判断 (主体判断) を加える。このとき、表現主体の認識構造が形成されこの認識構造を言語規範に基づいて表現する。また、この逆に、言語規範に基づいて言語表現から認識構造を捉え、その認識構造から対象のあり方を推論することにより言語理解が行なわれる。

認識構造は、表層構文構造をもとに構成されているが、客体認識 (格フレーム・深層格) と主体判断 (テンス・アスペクト・様相情報) に分離されている。認識構造は、言語間で共通の枠組みを構築しつつ、言語依存の構造も残した設計となっている。それぞれの言語の傾向により客体認識にはずれが生じるが、主体判断は言語間で比較的共通していると考えられる。

変換の際には、これらを分離して客体認識部分は格パターン変換を行ない、主体判断部分はそのまま英語側に渡し、英語の客体認識部分と融合する。

図1は、日本語文「昨日彼はみかんを食べた。」の認識構造である。

図2は、英文'She ate a cake.'の認識構造である。

```
|- 認識世界
    |- 客体認識
    |   |- 動的属性
    |   |   |- 概念 -- < 食べる >
    |   |- が格
    |   |   |- 対象
    |   |   |   |- 概念 -- < 彼 >
    |   |   |- 深層格 -- < 主体 >
    |   |   |- 実体把握 -- < 普遍性／特殊性 >
    |   |- を格
    |   |   |- 対象
    |   |   |   |- 概念 -- < みかん >
    |   |   |- 深層格 -- < 対象 >
    |   |   |- 実体把握 -- < 指向対象／動的目標 >
    |   |- φ格
    |   |   |- 概念 -- < 昨日 >
    |   |   |- 深層格 -- < 時間 >
    |- 主体判断
        |- 判断 -- < 肯定 >
        |- 時制 -- < 過去 >
        |- 態 -- < 能動 >
        |- 様相 -- < $nil >
        |- 相 -- < $nil >
```

図 1: 日本語認識構造

```
|- 認識世界
    |- 客体認識
    |   |- 動的属性
    |   |   |- 概念 -- < eat >
    |   |- 対象_Agent
    |   |   |- 概念 -- < she >
    |   |- 対象_Object
    |   |   |- 把握 -- < 不特定 >
    |   |   |- 概念 -- < cake >
    |- 主体判断
        |- 判断 -- < 肯定 >
        |- 時制 -- < 過去 >
        |- 態 -- < 能動 >
        |- 様相 -- < $nil >
        |- 相 -- < $nil >
```

図 2: 英語認識構造

5 非線形パターン

表 1 は、日英対訳データ「彼が失敗したので私は心が痛む。」⇔'It ails me greately that he failed.' から作成したパターンである。

表 1: 非線形パターンの例

日本語パターン	CL1 ので N2 は 心が 痛む
英語パターン	it ail N2.obj that CL1

※ N2.obj の部分は「名詞を目的格に変形する」意を表す。

このパターンでは、構文要素 (語、句、節) が汎化 (変数化) される。変数は、対応付けのために番号が振られている。汎化のレベルは部分によって異なる。半自動的な変数化の後、内省的に最適な汎化レベルに調整する。

6 線形パターン

線形パターンとしては、単文パターンを用いる。これには、格パターン辞書 (日本語語彙大系 5 構文体系 [3]) を利用する。認識構造の客体認識をもとにデータベースから検索を行ない、その情報を利用する。表 2 は格パターン辞書のレコードの例である。

表 2: 格パターンの例

フィールド	(例)
表記	味わう
日本語パターン	N1 が N2 を味わう
英語パターン	N1 experience N2
名詞意味属性	N1 4 人, N2 1259 苦しみ 1270 心配 1380 苦心 1262 悲しみ 1239 感覚

7 日英翻訳システムの概要

図 3 は、日英翻訳システムの概要を示したものである。

入力された入力文は日本語パーザ [4][5][6] による構文解析後、認識構造に変換される。非線形パターンが適用できる場合には、非線形変換モジュールで処理する。ここで、データベースから検索された非線形パターン

を利用して変換する。非線形パターンに部分的に単文が含まれている場合にはその単文の箇所については線形変換モジュールを利用して変換する。非線形パターンが適用できない場合には、線形変換モジュールによって要素合成的な変換を行なう。単文の場合には、線形変換モジュールのみ用いる。

非線形部処理モジュールでは、重文・複文レベルの日英文型パターンを利用して非線形な変換を行なう。

線形部処理モジュールでは、格パターン辞書(日本語語彙大系 5 構文体系)を利用して単文単位の変換を行なう。

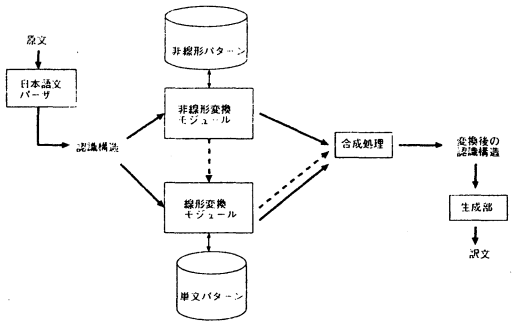

図 3: 日英翻訳システムの概要

8 非線形パターンのマッチング

ここでは、入力と非線形パターンのマッチングの流れを示す(図 4)。

入力された文は、形態素解析、構文解析を通じて構造化される。パターン側は形態素解析されている。パターン側から単語もしくは、変数化された構文要素ごとに入力とのマッチングを試みる。入力側は単語レベルでマッチングが行なえない場合には順次より大きな構文要素にさかのぼってマッチングを試みる。そのようにして最後まで成功すればパターンとのマッチング成功とする。

9 変換の具体例

ここでは、実際の変換の流れを具体的に示す(図 5)。入力文「彼が入学試験に落ちたので私は心が痛む。」を構文解析する。まず、非線形パターンとのマッチングを試みる。ここで、データベースから「CL1 ので N2 は 心が 痛む」⇔「it ail N2.obj that CL1」が得られる。変数 CL1 に相当する「彼が入学試験に落ちる」の部分は線形変換(単文変換)を行なう。格パターン辞書の「N1 が N2 に 落ちる」⇔「N1 fail in N2」{N1 … 3 主体, N2 … 1426 試験} がそれとマッチするのでこれを利用して変換する。結果として、「It ails me that he failed in the entrance exam.」となる。

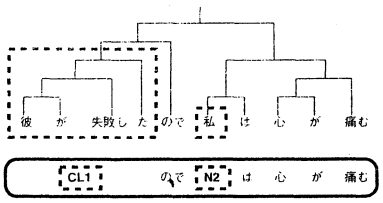

図 4: 非線形パターンのマッチング

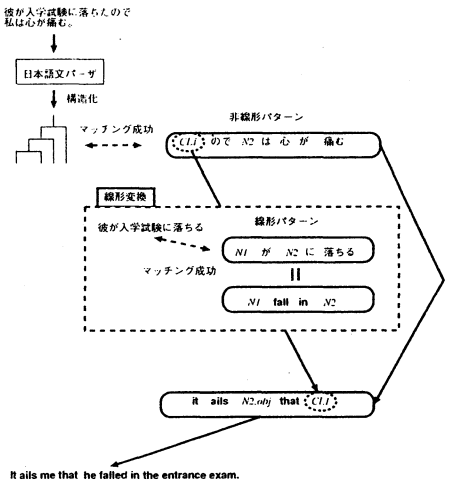

図 5: 変換の具体例

10 おわりに

本稿では、線形・非線形融合型日英変換方式を提案し、その有効性について論じた。本方式では、従来の単純な要素合成方式では適切な翻訳結果が得られないような場合にも品質の良い訳文が得られることを示した。

効果的な非線形パターンの準備方法については今後検討が必要である。

11 謝辞

この研究は、科学技術振興事業団 (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 [7]。

参考文献

[1] 時枝誠記:日本文法 口語篇、岩波全書 (1950).

[2] 三浦つとむ:日本語とはどういう言語か、講談社学術文庫 (1976).

[3] 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).

[4] 沼崎、田中:SGLR:逐次型一般化 LR パーザの Prolog による実現、情報処理学会論文誌、vol.32、No.3、pp.396 ~ 403(1991).

[5] 五百川、宮崎:痕跡処理のための逐次型一般化 LR パーザ SGLR の拡張、言語処理学会第 4 回年次発表論文集、pp.314 ~ 317(1998).

[6] 薮、藤石、宮崎: 表現構造と話者の対象認識構造を抽出する日本語文パーザの試作、言語処理学会第 3 回年次発表論文集、pp.205 ~ 208(1997).

[7] 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).