線形・非線形融合型日英構造変換のためのパターン記述形式

川辺 諭† 武本 裕‡ 宮崎正弘‡

†科学技術振興事業団 ‡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、節の統語レベルで汎化作業を行い、汎用性の高い日英文型パターンを作成する。日英/英日間の翻訳処理において、入力文に適合するパターンデータを検索し、得られたパターンに基づいて出力文を生成することで、原言語表現に対応する、こなれた目的言語表現を出力することが可能になる。本稿では被覆率の高い日英文型パターンの記述形式を提案し、その有効性について論じる。

2 研究の目的

日英翻訳においては、日本語、英語文の間で単語や統語構造の対訳関係が決定できない“非線形”な部位があり、原言語中の単語を単純に置き換えただけではこなれた目的言語表現を得ることはできない。

この問題点を解決するために本稿では、以下の手法を提案する。

1. 非線形に対応する部位を含む日英対訳文(重文、複文)のパターン化を行い、日英文型パターンデータとして準備しておく。
2. 翻訳処理時は、原言語入力文の形態素解析、構文解析を行う。
3. 解析された入力文の構造を利用し、適合する日英文型パターンを検索する。
4. マッチングしたパターンを参考に出力文の大域的な構造を決定する。
5. マッチングしなかった差分について、さらにパターン検索を繰り返す。
6. 検索に成功した場合は、パターンを参考にして、すでに準備されている大域構造に、局所構造を埋め込む。
7. パターンにマッチしなかった差分は文型パターンを使うなどして合成する。

図1は試作中の日英翻訳処理システムの概要である。

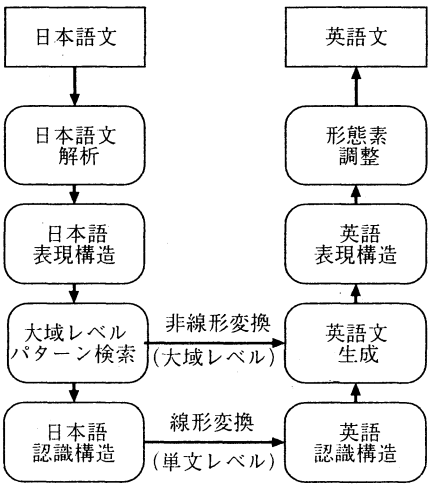

図1: 日英翻訳処理システムの概要

3 日英文型パターンの作成

言語データから日英文型パターンを作成する手法についてのべる。

3.1 日英文の汎化

1. 語レベル汎化
   名詞(N)、動詞(V)、形容詞(AJ)など、日英間で線形に対訳関係が決まる自立語を変数化する。日本語側の変数に昇順で番号を振り、英語側の対応する要素(品詞が異なることがある)に、同じ番号を与える。それぞれの要素に付随する時制、相、様相などの屈折情報は、変数に ".”(ドット)をはさんで後置する。
2. 句レベル汎化
   線形要素とみなすことができる名詞句(NP)と動詞句(VP)を変数化する。番号は主辞のものを引き継ぐ。動詞句内の助詞や助動詞の持つ屈折などの情報に関して、文型パターン選択の際に必要となる情報を、後置して記述する。
3. 節レベル汎化
   線形要素とみなすことができる節(CL)を変数化する。番号は述部のものを引き継ぐ。句レベル汎化同様、文型パターン選択に必須となる情報を、CLに後置して記述する。

3.2 日英文型パターンの例

日本語文、英語文の汎化作業と、文型パターンの例を示す。

```
JL0 : 彼が勝ったので(私は)驚いている
JL1 : N1が V2.た ので (N3は) V4.ている
JL2 : N1が VP2 ので (N3は) VP4
JL3 : CL2 ので CL4
EL0 : I am astonished that he has won.
EL1 : N3 V4.pass  that N1 V2.pf
EL2 : N3 VP4 that N1 VP2
EL3 : CL4 that CL2
*JP : CL2 ので CL4
*EP : CL4 that CL2
```

図1: 日英文型パターンの例(1)

図1において、JL0とEL0は日本語、英語原文であり、J(E)L1、J(E)L2、J(E)L3は語、句、節レベルの汎化を行ったデータである。EL1の動詞Vに後置されている ".pass" や ".pf" が、時制、相情報を表している。

日本語原文JL0の助詞「た」や助動詞「ている」は、字面だけでは完了、継続などといった相情報を限定できない。このような情報はJL1における ".た" や ".ている" のように、字面表記で残す。

最終的には図1の*JP、*EPの内容が、日英文型パターンとして抽出される。

```
JL0 : 彼が失敗したので私は心が痛む
JL1 : N1が V2.た ので N3は N4が V5
JL2 : N1が VP2 ので N3は VP5
JL3 : CL2 ので CL5
EL0 : it ails me greatly that he failed.
EL1 : it V5 N3.obj AV6 that N1 V2.ed
EL2 : it VP5 that N1 VP2
EL3 : CL5 that CL2
*JP : CL2 ので N3は 心が 痛む
*EP : it ail N3.obj that CL2
```

図2: 日英文型パターンの例(2)

図2は図1と同様に語、句、節レベルの汎化を行ったものである。EL1の ".obj" と ".ed" は、それぞれ "目的格への変換" と "動詞の過去形屈折" を表す。

この例では日本語側が「私は心が痛む」の節にハガ構文を、英語側は形式主語 "it" の構文を用いており、表面的な統語構造が単純には対応していない。また、英語側表現は "that he failed"(彼が失敗したこと)を主語としており、日英表現間において表現対象の捉え方が本質的に異なっている。このように言語の特性に強く依存する部分は、上位レベルまで汎化せず、字面や語、句のラベルとしてパターン中に残す。

最終的に得られるパターンが図2の*JP、*EPである。日本語、英語のパターン中に、「~は心が痛む」"it ails ~ that" といった、構文を特徴づける部分が残されている。このように日英文型パターンは、字面、語変数(N)、節変数(CL)が混在したものになる。

3.3 人手で準備するパターン

日英文型パターンを網羅的に準備し被覆率を向上するために、準備された言語データに含まれないいくつかのパターンに関して、作業者の言語学的な内省をもとに、直接人手で準備するパターンもある。図3は英語 "so-that" 構文の文型パターンである(語、句、節レベルの汎化の様子は省略)。

■ No.1
JP : N1は (大変*) AJP2で CL3
EP : N1 be (so) AJP2 that CL3
■ No.2
JP : N1は (大変*) AVP2 VP3 で CL4
EP : N1 VP3 (so) AVP2 that CL4

図3：“so-that” 構文の文型パターン

英語 “so-that” 構文は、日本語側が「～は大変～なので～」と訳出されることが多いが、強意を示す副詞「大変」は、「とても」など同等の副詞に置き換えることができる。パターン中でこのことを示すために、“(大変*)” と記述している。

また、「大変」で強調される内容が述部の形容詞であるか、述部の動詞であるかによって、英語側で強調される部位や文型の構造が異なるため、2通りのパターンを準備する必要がある。

図4は “no-sooner-than” 構文の文型パターンである。

■ No.3
J0:彼が家を出るとすぐ雨が降り始めた
E0:He had no sooner left home
than it began to rain.
JP:N1が VP.-た と すぐ N3が VP4.た
EP:N1 had no sooner VP2.pp than N3 VP4.past
EP:No sooner had N1 VP2.pp than N3 VP4.past
■ No.4
J0:彼はそれを終えるとすぐ読書を始めた
E0:No sooner had he done it
than he began reading.
JP:N1は VP2.-た と すぐ VP3.た
EP:N1 had no sooner VP2.pp than N1 VP3.past
EP:No sooner had N1 VP2.pp than N1 VP3.past

図4：“no-sooner-thant” 構文の文型パターン

図4の2つの日英文型パターンは、英文パターンEPの後半部分の主語が異なっている。これは、No.4ではJP文頭の “は格” が後件の主格を兼ねるためである。

日英翻訳処理システムの被覆率を向上するために、いくつか日英文型パターンに関して、ここに示したように人手による作業でパターンを補う必要がある。

4 処理例

以下は日本語文「彼女が怪我をしたので彼は心が痛む」に関して、日英文型パターンを検索し、英語文を生成する様子である。

■ 入力文
彼女が 怪我を し た ので 彼は 心が 痛む
■ 適合した日英文型パターン
JP : CL1 ので N2は 心が 痛む
EP : it ail N2.obj that CL1
■ 大域構造決定
-> it ail N2.obj that CL1
■ 単文生成
J_CL1 = 彼女が 怪我を した
E_CL1 = she was injured
■ 変数と単文の埋め込み
-> it ail he.obj that she was injured
■ 形態素調整
-> it ails him that she was injured

5 おわりに

日英翻訳において、単語同士が単純に対応しない非線形対応部を持つ日英対訳データから、日英文型パターンデータを作成する手法を提案した。入力文に適合する日英文型パターンを利用することで、品質の良いこなれた出力文が得られるようになる。

6 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。

参考文献

[1] 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).