高品質な日英単文用例翻訳システムの枠組み

藤田 和寛　　宮崎 正弘
新潟大学大学院自然科学研究科

1　はじめに

要素合成方式の機械翻訳では、入力文を単文の組合せとなるように分解し、それぞれを翻訳したものを再び組合せることによって、出力となる文を生成する。また、複文・重文を扱うパターン翻訳でも、日本語側パターン中の変数が節･動詞句などの場合、単文相当の部分を翻訳し、その結果を英語側の対応する部分に埋め込む必要がある。しかし、高品質な機械翻訳を行うためには、各単文がそれぞれ適切に翻訳される必要があるにもかかわらず、現在商用化されている日英機械翻訳システムはこなれた日本語の単文を自然な英文に確実に翻訳するだけの能力を有していない。

本稿では、高品質な日英単文機械翻訳システムに必要となる手法について検討し、全体的な翻訳システムの枠組みについて論じる。

2　主体表現の翻訳

2.1　主体表現

時枝誠記の言語過程説およびそれに基づく三浦つとむの日本語文法体系によれば、言語の意味は表現自体が持っている対象と認識との間の客観的な関係であり、言語表現を行う表現主体である人間は、対象を概念として認識し（客体認識）、その認識に対してある種の判断（主体認識）を加える[1]。

主体表現の表層的表現は多様であるが、その内容は話者の主観的判断や感情に関する情報であり、ある程度のパターンに限定される。そのため同様な情報を類別し、素性として扱うことが可能であり有効となる。

主体表現の類別のために、同一単文内に同時に存在しない属性の排他性を考慮したグルーピングを行い、ノード数 200 余りの主体表現属性体系を用意した。各属性は、英文生成プログラム作成上の便宜のため、判断・態・時制・相・様相・実体把握からなる大分類と、それをさらに中分類・小分類・細分類に細かく分類した、4 階層の属性名を付加してある（図 1）。

判断	肯定	単純断定		0110	だ(った)?	肯定判断辞
態	利益授受	受益		1410	てくれ[るた]	
時制	過去			2100	た	既定判断辞
相	未開始			3100	ずにい[るた]	
様相	実現性	難易	困難	4721	にく(い|かった)	形容詞型接尾語

図 1：主体判断属性（一例）

2.2　主体表現の抽出

主体表現は主に助詞・助動詞・感動詞・接続詞・陳述副詞によって表される。助詞などの客体表現の構成要素となるもの以外の表現については、データベースに登録されている日本語表現に対応する部分を入力文から直接抽出する。効率よく正確に抽出するために、入力文を形態素解析し、形態素の区切りと品詞情報を利用する。

形態素解析された文は、後ろから 1 形態素ずつ増やしながらデータベースと比較を行う。表現によってはとりうる品詞が決まっているため、同時に品詞情報も比較する。マッチした最も長い表現について、属性を記憶し、入力文から該当部分を削除する（図 2）。一つの表現に複数の属性がある場合には、全ての属性を記憶する。

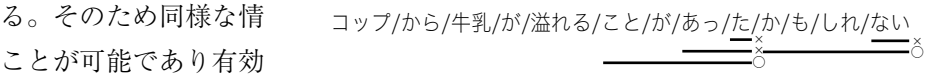

図 2：主体表現の抽出方法

2.3　実体把握の表現

客体表現を構成する要素の中には、日本語ならは助詞、英語ならば冠詞、前置詞といった実体に対する話者のとらえ方、つまり主体表現を表す語も含まれる。日本語の助詞は実体に対する話者のとらえ方を表しており、副助詞・係助詞は実体に対するよりきめ細かいとらえ方を表す。例えば「鳥が飛ぶ」と「鳥は飛ぶ」では、個別性を表す表現と普遍性を表す表現という違いがある。また、「～も～も」のように並列を表す表現や、「～さえ」のような強調を表す表現なども客体表現の構成要素となる。

このような表現を表1に示すような実体把握の素性として分離し、格要素が持つ情報として英文生成部に渡す。

表1：実体把握の素性（一例）

素性		例
並立	単純列挙	～も～も，～と～と，……
強調		さえ，すら，こそ，……
例	例示	など，なんか，なんて，……
添加		さえ，すら，でも，……
限定	個別性	が

2.4　主体表現の受け渡し

以上の処理によって、文中にどの主体表現が存在するかを抽出することができる。多くの場合、各表現の存在の有無のみの情報でも十分に英文を生成することができる。しかし、場合によってはさらに詳しい情報を英文生成部に渡す必要がある。

主体表現の順序

複数の主体表現の現れる順番によって意味に違いが現れることがある。例えば「コップから牛乳が溢れることがあったかもしれない。」と「コップから牛乳が溢れるかもしれないことがあった。」では意味合いが異なる。これは「ことがあった」と「かもしれない」がかかる範囲の違いによるものである。この違いを英文にも反映されるためには、表現が出現する順番の情報を英文生成部に渡す必要がある。

「た」の扱い

過去・完了を表す「た」は複数回現れることがある。複数の主体表現を表す部分それぞれが現在と過去の区別を行いうるが、それぞれの時制の組合せが違うことによる意味の違いは、人間でも区別するのは難しい。そこで、過去・完了を表す「た」については、本動詞部分とその他の部分の2つの区別のみ行うことにする。具体的には、終止形でない動詞の直後にある主体表現に「た」がついていた場合には時制“過去”を記憶し、その他の主体表現に一つでも「た」がついていた場合には“助過去”を記憶する（図3）。

コップから牛乳が溢れ**た**かもしれなかっ**た**
（過去）（助過去）

コップから牛乳が溢れるかもしれなかっ**た**
（現在）（助過去）

コップから牛乳が溢れ**た**かもしれない
（過去）（φ）

図3：「た」の属性

以上の操作によって得られた属性について、未来判定、「れる・られる」様相判定、その他の格変換を行う。この結果を英文生成部に渡し、原形となる英語表現に助動詞や前置詞の挿入、種々の語の屈折（変形）を行うことになる。

3　客体表現の翻訳

3.1　客体表現の翻訳手法

客体表現は、それぞれの言語に固有な様々な表現を持ち、多様な内容を持つ。機械翻訳は大きくルールベースの手法と用例ベースの手法とに分けられるが、客体表現の表す微妙なニュアンスの違いを表現するためには、膨大な客体表現の表現手法を人手によってルール化するよりも、用例ベースの手法を主体にするのが有利である。

3.2　用例の準備

用例対は、準備の自動化を見越した、最低限の変換とタグ付けを行う。用例は、あらかじめ修飾語句

が無い状態にしておく。日本語文は形態素解析して分かち書きされて品詞タグの付いた状態にし、動詞は終止形にする。英語文は、日本語文の格要素に対応する単語を日英単語辞書を用いて検索し、それぞれに格のタグを付ける。また、英文生成部が処理しやすいように、主部（NP）と述部（VP）を区別するためのセパレータを挿入し、動詞は原形にする（図 4）。

```
名詞:杯 格助詞:が 名詞:酒 格助詞:で 動詞:溢れる 句点:。
#The glass_が|be filled to the brim with liquor_で
名詞:彼 係助詞:は サ変型名詞:希望 格助詞:に 動詞:溢れる 既定判断辞:て 形式動詞:い 既定判断辞:た 句点:。
#He_は|be brimming over with hope_に
```

図 4：用例対の例

検索方法指示記号

用例の中には、極端に汎用性の高い構文をもったものや、限定性の高い表現があるため、それらの用例には指示記号を付加し、用例のマッチ方法を変更する。例えば「～は～より良い」という表現では、比較対象として使える語句には制限がないと考えて良い。このような用例には、用例の検索にシソーラスを使わず、どのような語句でも使ってよいことを指示する。逆に「涙が溢れる」の場合、英語文は“The eyes is filled with tears”のように、日本語では省略されている“eyes”が表現に現れざるを得ないが、ここで“涙”以外のものが“tears”の部分に当てはめられることを防ぐ必要がある。このような用例に対応するため、用例の検索時に、シソーラス中の同じカテゴリにのみマッチさせることを指示する記号と、完全に同じ単語のみとマッチさせることを指示する記号を導入する。

3.3　パターンの利用

特定の表現では、用例を用いるよりも、要素を記号化したパターンを用いた方が、より効率的に細かな意味合いを対応づけられると期待できる。例えば「象は鼻が長い」のように 1 単位文中にハ格とガ格が同時に存在する文とその類似表現では、それぞれに対して適切となる英語表現は、定／不定や総称／指示を示す構造がそれぞれ異なる[2]が、それぞれの構造は別の名詞や形容詞でも同じになると考えられる。従って、表 2 のようなパターンを用意することによって、細微な表現の違いまで網羅的に用例を準備することなく訳し分けることが可能になる。

表 2：パターンの例[3]

日本語表現	英語表現
N1 は N2 が *adj* （象は鼻が長い）	N1s have *adj* N2s （Elephants have long trunks）
N1 が N2 が *adj* （象が鼻が長い）	The N1 has a *adj* N2 （The elephant has a long trunk）
N1 は N2 は *adj* （象は鼻は長い）	N1's N2s are *adj* （Elephants' trunks are long）
N1 が N2 は *adj* （象が鼻は長い）	The N2 of the N1 is *adj* （The trunk of the elephant is long）

3.4　用例の選択

用例ベースの手法を用いる場合、いかに適切な表現を持った用例を選択できるかが要点となる。本枠組みでは、用例となる文にはほぼ格要素と動詞・形容詞・形容動詞のみが含まれることになるので、比較的シンプルな方法を用いることができる。

従来、入力文と用例に含まれる格要素同士のシソーラス中の距離を点数化する方法がよく用いられてきた。しかし、分類の細かさがシソーラス中で一定になっていないために、ノード間の距離の違いと実際の意味的な距離に隔たりが出る場合がある。また、用例との距離の許容範囲を広くとれば、用例とマッチする可能性は高くなるが、その分不適切な用例とマッチする可能性も高くなってしまう。従って不適切な用例を選択する可能性を低く抑えるためには、シソーラス中の距離が近いものに限定するようなマッチ方法を用いなければならない。

用例の検索では、まず動詞・形容詞・形容動詞が等しく、格の要素が等しいものを選ぶ。その中から格となる名詞から連想される意味の近さを検討する。例えば、「コップから牛乳が溢れる」という入力文では、「コップ」と「牛乳」は共に“飲食”という概念に関連している[4]。これは「杯から酒が溢れる」という同様に“飲食”の概念を持った用例と

一致する（図 5）。

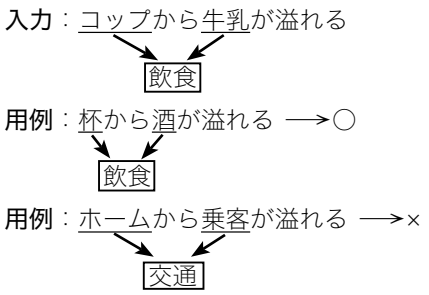

図 5：連想による用例検索

検索に連想を利用できない場合には、それぞれの格の名詞同士がシソーラス[5] 中で同じノードまたは兄弟のノードにあるものを採用する（図 6）。用例が見つからない場合に、パターンを利用した翻訳を試みる。

```
|        | |[-] 853嗜好品等
|        |   |-854果物
|        |   |-855菓子
|        |   |[-] 856飲物･たばこ
|        |     |[-] 857飲物
|        |     | |-858茶
|        |     | |-859コーヒー・ジュース
|        |     | |-860ミルク
|        |     | |-861酒
|        |     |-862たばこ    「牛乳」とマッチする範囲
|        |-863建造物 [+]
|        |-893道具 [+]
|        |-962機械 [+]
|        |-986乗り物 [+]
|-1000抽象 [+]
```

図 6：シソーラスで名詞がマッチする範囲

4　まとめ

本枠組みで単文の翻訳を行う手順をまとめると以下のようになる（図 7）。入力文を形態素解析し、客体表現の構成要素とならない部分の主体表現を抽出する。得られた主体表現の情報を元に未来判定、「れる・られる」様相判定、格変換を行う。さらに残りの客体表現から実体把握を表す表現を抽出し、格変換する。以上の処理によって得られた客体表現を用いて、連想による用例検索、シソーラスの兄弟関係を用いた用例検索、パターン検索の順に検索を行い、基本となる英語文を生成する。この文を英文生成部に主体表現情報と共に渡し、最終的な出力を生成する。

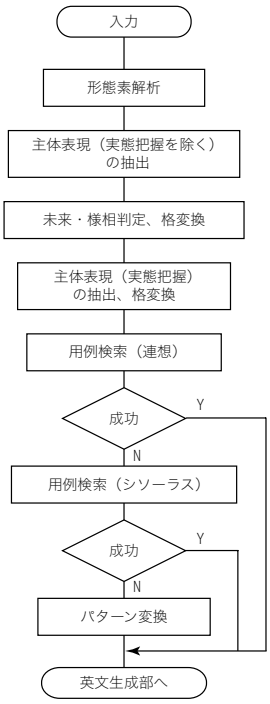

図 7：翻訳の流れ

参考文献

[1] 三浦つとむ（1976）日本語はどういう言語か，講談社学術文庫

[2] Francis Bond, Kentaro Ogura（1998）Reference in Japanese-English Machine Translation, *Machine Translation,* **13**, 107–134

[3] 白井諭，Francis Bond，野沢弥生，佐々木富子，上田洋美（1999）入力文と結合価パターン対辞書の照合に関する一手法, 言語処理学会第 5 回年次大会, C1–2, 80–83

[4] 森田陽介, 宮崎正弘（2006）連想型多次元シソーラスとその意味解析への適応性，言語処理学会第 12 回年次大会，A4–2, in press

[5] 池原悟，宮崎正弘，白井諭，横尾明男，中岩浩巳，小倉健太郎，大山芳史，林良彦（1997）日本語語彙大系，岩波書店