結合価パターンを用いた動詞句の翻訳可能性の調査

石上 真理子　徳久 雅人　村上 仁一　池原 悟
鳥取大学工学部知能情報工学科
{isigami,tokuhisa,murakami,ikehara}@ike.tottori-u.ac.jp

1　はじめに

言語表現の非線形性の問題に対処する機械翻訳の方式として，文型パターンを用いる方式が提案されている [1]．この方式では，線形要素は変数で表し，変数に対する訳出表現を組み合わせて全体の英文を作成する．そこで,「線形要素の翻訳が，代入された日本語だけで可能である」という仮説を立てている．

本稿では，変数の中でも動詞句の翻訳に焦点を当て，動詞句の翻訳可能性について調査する．調査のねらいは，次の 2 点である．

1．局所翻訳の可能性：動詞句の翻訳において，上述の仮説の検証

2．動詞句の非線形性：動詞句に含まれる非線形な部分の検証

まず，1 については，一般の翻訳ソフトウエアを用いて検証する．本稿では，既存の 2 つのソフトウエア（以下，システム 1，システム 2 と称す）を用いて訳出の可能性を調べる．

次に，2 については，動詞句に含まれる非線形性の種類が問題となる．動詞句の非線形性は，格要素と述語の観点からは，結合価パターン [3] が網羅的に示している．本稿では，まず，非線形性のある格要素と述語の関係が結合価パターンだけでどの程度カバーできるかを検証し，その他の非線形部分についても調査する．

2　翻訳プロトタイプ「queen」

2.1　結合価パターン

結合価パターンは，用言と格要素（名詞＋格助詞）の意味的関係を記述している．[3] は，約 15,000 件の日英文型パターン対を集収している．例を以下に示す．

$N1$(3 主体) が $N2$(3 主体) を 援助する	$N1$ help $N2$
$N1$(838 食料) が 口に 合う	$N1$ taste good

2.2　翻訳手順

翻訳アルゴリズムを例とともに以下に示す．

1. 日本語の動詞句の入力
   入力する動詞句の例を以下に示す．
   ・パソコンを使う。
2. 結合価パターンとの照合
   入力した動詞句と結合価パターンとの照合を行う時の条件を以下に示す．
   (a) 格要素の省略，格要素の順序変更，修飾語句の挿入，任意追加の格要素を許可する．
   (b) 意味属性制約，必修の述語付属語の条件を必須とする．
   入力する動詞句の例と適合したパターンを以下に示す．
   ・$N1$ employ $N2$ as $N3$
   ・$N1$ use $N2$ as $N3$
   ・$N1$ spend $N2$ on/for $N3$
3. 照合の結果の選択
   慣用句表現を優先的に選択し，最も広範囲に入力文と一致するパターンの選択を行う．適合したパターンから，選択されたパターンを以下に示す．
   ・$N1$ use $N2$ as $N3$
4. 名詞変数への訳語の挿入
   既存の日英辞書引きプログラムを用い，入力動詞句と適合する変数に訳語を代入する．適合しない残った変数については，目的語ならば文の骨格を左右するため残す．しかし，補語ならば骨格を左右しないため消去する．選択されたパターンの名詞変数への挿入を以下に示す．
   ・$N2$=パソコン→ personal computer
   また,「as $N3$」については，入力動詞句と適合していなく，補語であるため消去する．
5. 英語の動詞句の形成
   主語に該当する変数を削除し，句を形成する．句を形成した例を以下に示す．
   ・use personal computer
6. 句形成した英文を出力する．

本プロトタイプでは，主動詞と格要素の訳出構造をとらえることが目的である．そこで，名詞訳語の選択，

任意の格要素（時間，場所，手段など）の訳語挿入，副詞の挿入，冠詞・数の問題は扱わない．

2.3 翻訳用知識ベース（結合価パターンの変形）

結合価パターンは，それぞれ一文で表される．本稿では，文ではなく動詞句を扱うため，2.2 節の「英語の動詞句の出力」で示した処理を施す．基本となる結合価パターンを変形した例を以下に示す．

（変更前）：$N1$ use $N2$ as $N3$

（変更後）：use $N2$ as $N3$

3 動詞句の翻訳実験

3.1 目的

本稿は，動詞句の局所翻訳の可能性および動詞句の非線形性を検証することを目的とする．そのため，queen，システム 1，システム 2 による動詞句のそれぞれの訳出を出力する．そして，訳出品質の評価値を付与する．なお，システム 1 では，非線形性のある格要素と述語の関係は [3] でカバーしており，さらに線形要素についてもカバーしている．

3.2 翻訳対象

[1] では，重文・複文の 15 万文対の日英対訳コーパスから文型パターンを作成した．その作成過程において，対応関係の見出された動詞句が約 7 万件存在する．コーパスの例を以下に示す．下線部が翻訳対象である．

日本文：大勢の前で話すときは多くの人があがるものだ。

日本語パターン：/y$\underline{VP1}$^rentai!ときは/tcfk$NP2$が/cf$V3$^rentai!ものだ。

対訳英語文：Many people have stage fright when they have to give a speech in front of a large audience.

対訳英語パターン：$NP2$ $V3$ when $N3$ have to $\underline{VP1}$.

日本語パターンにおいて，動詞句 ($VP1$) になっているところは，日本語文では，「大勢の前で話す」に相当し，対訳英語では，「give a speech in front of a large audience」に相当する．つまり，「大勢の前で話す」という動詞句と，「give a speech in front of a large audience」は対訳関係にあると言える．

本稿では，この対訳コーパスから，ランダムに取り出した 325 句対の動詞句を翻訳対象とする．ただし，英語句の構成が 2 単語から 14 単語までのものを，それぞれ 25 句ずつ取り出した．

3.3 訳文の評価

- 訳文の評価値と評価基準

  訳文を検証する評価値とその評価基準を以下に示す．

  評価◎： 主動詞と主名詞が理想解と完全一致．

  評価○： 理想解と異なる構文だが，意味は合っている．

  評価△： 基本的な構造は良いが，出力結果の字面の一部を変形する必要がある．

  評価×： 訳出が間違っている．

  評価ー： 出力パターンが一つもない．

  ただし，線形翻訳で追加できる前置詞句は問題外とする．評価の例は，3.4 節に示す．

  なお，システム 1，システム 2 は動詞句の出力にするため，主動詞を命令形にして入力する．

- 評価方法

  評価方法として再現率・適合率を用いる．定義を以下に示す．

  再現率 R：全入力事例数における，出力パターンが一つでもある事例数 N の割合

  適合率 P：N における，総合評価が◎または○である事例数（スコア）の割合

3.4 実験結果

queen，システム 1，システム 2 のそれぞれの翻訳結果を表 1 に，そして再現率 R と適合率 P を表 2 に示す．

表 1: 3 つのシステムの翻訳結果

	評価（個数）				
	◎	○	△	×	ー
queen	86	135	36	34	34
システム 1	74	156	24	69	0
システム 2	84	162	25	54	0

queen の評価の例を以下に示す．

- 評価◎の例

  入力動詞句：パソコンを使う。

  理想解：use personal computer

  適合パターン：$N1$ use $N2$ as $N3$

  出力結果：use personal computer

  代入値：$N2$=パソコン→ personal computer

- 評価○の例

  入力動詞句：病気に成る。

  理想解：result in her illness

  適合パターン：$N1$ fall sick

  出力結果：fall sick

- 評価△の例
  入力動詞句：都会に移る。
  理想解：moved to the city
  適合パターン：$N1$ move from $N2$ to $N3$
  出力結果：move from ($N2$) to city
  代入値：$N3$=都会→ city
  下線部は，出力結果を変更しなければならない字面の部分である．
- 評価×の例
  入力動詞句：足に怪我をする。
  理想解：hurt my leg
  適合パターン：$N1$ make $N2$ $N3$
  出力結果：make accident foot
  代入値：$N3$=足→ foot，$N2$=怪我→ accident
- 評価－の例
  入力動詞句：頭から追い出す。
  理想解：get out of our heads
  適合パターン：なし

表 2: queen，システム 1，システム 2 の精度

	再現率 R	適合率 P
queen	89.5%(291/325)	75.9%(221/291)
システム 1	100%(325/325)	70.1%(230/325)
システム 2	100%(325/325)	75.7%(246/325)

表 2 から，いずれのシステムも，再現率 R，適合率 P ともに比較的高い値を得た．よって，主動詞と主名詞の構造に関しては，代入された日本語だけからの翻訳の可能性が 70%程度であることが分かった．

4 考察

4.1 局所翻訳の可能性の検証

構成単語数ごとのスコア（0 から 1 に正規化したもの）を図 1 に示す．

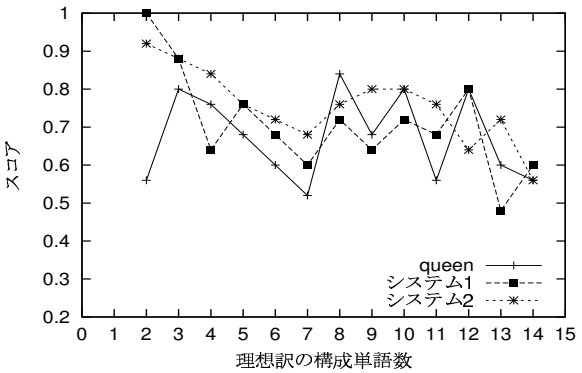

図 1: 各単語数でのスコア

図 1 より，構成単語数が 2 単語から 7 単語に近付くにつれて，スコアが減少する傾向がある．しかし，7 単語を過ぎると，0.5 から 0.8 までの範囲に収まっている．

これは，主動詞と主名詞の構造を中心に評価したためと考えられる．今後の課題に，その他の部分の訳出構造を検討することがあげられる．

4.2 動詞句の非線形性の検証

queen の翻訳実験において，主動詞と主名詞以外の訳出について，評価時に分かったことを分析する．

分析の結果，動詞句の翻訳においては，格要素と動詞の関係だけでは判断できない非線形性と思われるものがあった．その具体例を以下に示す．

1. 格要素が複雑である場合
   例）
   入力動詞句：異なる字体の文字 や グラフィックス までも 出力できる。
   理想解：are capable of producing characters in different type fonts and even graphic output
   この場合は，格要素が複雑であるため,「A を出力できる = are capable of producing A」というパターンがあったとしても，訳出は難しい．
2. 入力文に目的語がない場合
   例）
   入力動詞句：型紙に合わせる。
   理想解：fit the pattern
   適合パターン：$N1$ match $N2$ with $N3$
   出力結果：match ($N2$) with pattern
   英語の動詞が他動詞であるにも関わらず，入力文に該当格要素が存在しなかった．つまり,「合わせる」の目的語の情報がない．これは，英文生成時の問題と言うこともできる．
3. 再帰代名詞の変数が残っている場合
   例）
   入力動詞句：人間に化ける。
   理想解：take the shape of a human being
   適合パターン：$N1$ disguise $N1$-self as $N2$
   出力結果：disguise ($N1$)-self as human being
   代入値：$N2$=人間→ human being
   この場合は，格要素と述語の関係だけで決定することができない．また，この問題は，局所翻訳の可能性の問題でもある．つまり，$N1$-self の訳出は，主語の情報が必要である．動詞句の翻訳でも，主語を入力する必要がある．

1 章で述べたように，線形，非線形の区別は表現の部分と全体の関係を言うものであり，線形要素であってもその要素自体が非線形であることを意味しない．つまり，線形要素の内部構造は非線形であっても良い．今後は，結合価パターンで表現できない非線形構造について，別途パターン化することが必要と考えられる．

5　おわりに

4.1 節の局所翻訳の可能性の検証と，4.2 節の動詞句の非線形性の検証の結果より，主動詞と主名詞の構造に関して，文全体から見て線形要素とする動詞句については，代入された日本語だけからの翻訳の可能性が 70%程度であることが分かった．そして，動詞句の中の非線形性は，格要素と動詞の共起から生じるものの他に，副詞との共起，格要素自体の複雑さも問題になることが分かった．今後，より詳細に分析を進める．

参考文献

[1] 池原悟, 阿部さつき, 徳久雅人, 村上仁一：非線形な表現構造に着目した重文と復文の日英文型パターン化, 自然言語処理, Vol.11, No.3, pp69-95, 2004.

[2] 金出地真人, 徳久雅人, 池原悟, 村上仁一：結合価文法による動詞の訳語選択能力の評価, 自然言語処理, Vol.11, No.3, pp149-164, 2004.

[3] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦：日本語語彙大系, 岩波書店, 1997.

[4] 徳久雅人, 村上仁一, 池原悟：文型パターンパーサの試作, 言語処理学会第 10 回年次大会発表論文集, pp608-611, 2004.