# 結合価パターンを用いた日中機械翻訳方式の検討

楊 鵬　村上 仁一　徳久 雅人　池原 悟
鳥取大学 工学部 知能情報工学科
{s022061,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

## 1　はじめに

高品質な翻訳の実現を目指して，大規模な文型パターン辞書を用いた翻訳方式の研究が行われている [1]．日英翻訳では，すでに網羅的な結合価パターン辞書 [2] が開発され，単文の翻訳において，結合価パターンの方式が大変有効であることが報告されている．そこで，本研究では，日中翻訳におけるこの方式の有効性を評価するため，結合価パターン辞書を部分的に作成し，その効果を評価する．

ところで，結合価パターンによる方式は，語義数の多い用言の訳し分けにおいて特に効果が期待される方式である．通常，語義数の多い用言は，使用頻度も高いと考えられるので，本研究では，使用頻度の高い日本語結合価パターンを対象に日中結合価パターン辞書を作成し，使用する．

## 2　日中結合価パターンの部分試作

本研究では，すでに開発されている日英結合価パターン辞書の日本語パターンを対象に，対応する中国語結合価パターンを試作し使用する．

ところで、結合価パターンは，体言と用言の意味的な関係をパターン形式で表現したものであり，機械翻訳では，原言語のパターンと対応する目的言語のパターンを対にした結合価パターン対辞書が使用される．パターン対辞書の設計では，原言語のパターンに適合した入力文に対して，目的言語のパターンが一意に決定できるようにする必要がある．このため，参照する日英パターン辞書では，日本語パターンは，適合したすべての日本文が対応する英語パターンを用いて翻訳できるように，カバー範囲が決められている．従って，日中翻訳の場合は，この辞書に登録された日本語パターンは，必ずしも中国語のパターンに 1 対 1 に対応しない可能性がある．

そこで，本研究では，作成した日中結合価パターン辞書を用いた翻訳実験から，日中翻訳における日本語パターンの構成法上の問題点とそれを用いた結合価パターン方式の可能性を明らかにする．

### 2.1　日本語語彙大系の結合価パターン

本研究で使用する日本語語彙大系 [2] は，「構文体系」と「意味体系」から構成される．「構文体系」には，日本語の用言 6000 語に対して，一般文型 (11,500 件) と慣用表現文型 (3,300 件) の合わせて 14,800 件の結合価パターンが収録されている．「意味体系」には，日本語約 30 万語に対する意味属性を掲載されている．収録されている結合価パターンの例を以下に示す．なお，( ) 内の数値は意味属性である．

- 日本語結合価パターン：
  $N1$ "が" $N2$ "に" 無い
  $N1$ の意味属性：(*全ての意味属性)
  $N2$ の意味属性：(388 場所 533 具体物 1000 抽象)

### 2.2　中国語結合価パターンの作成方法

以下に示す 3 つのステップにより，日本語結合価パターンに対応する中国語結合価パターンを作成する．

1. 対象とする日本語結合価パターンの選択
   日英対訳コーパス（100 万件）の中から，日本語単文 27 万件 [3] を抽出する．次に，日本語語彙大系 [2] の結合価パターン辞書と照合し，使用頻度の高い日本語結合価パターン（上位 200 件）を選択する．
2. 中国語訳文の付与
   上記で選択された各日本語結合価パターンに対して，適合する日本文を 1 文選択し，対応する中国語訳文を付与する．
3. 中国語結合価パターンを作成
   上記２. で得られた日中対訳文を参考に，１. で選択した日本語結合価パターンに対応する中国語結合価パターンを作成する．

作成した結合価パターン対の例を以下に示す．

- 日本語結合価パターン：
  $N1$ "が" $N2$ "に" 無い
- 中国語結合価パターン：
  "在" $N2$ "里" 没有 $N1$
- 体言の意味属性：
  $N1$：(*全ての意味属性)
  $N2$：(388 場所 533 具体物 1000 抽象)

## 3　結合価パターンによる翻訳能力の評価

### 3.1　結合価パターンを用いた翻訳方法

翻訳手順を以下に示す．

1. 入力文の選択：
   作成した各日中結合価パターン（200 パターン）に対

して，日本語単文集 [3] から，その作成で参照した日本文と異なる例文 1 文（合計 200 文）を任意に選択し，試験文とする．試験文の例を図 1 に示す．

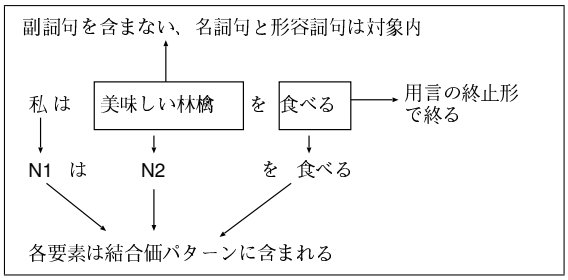


図 1　選択された日本語単文例

2. 中国語訳文の作成：

日中結合価パターン対を使用し，以下のようにして，各日本語入力文に対する中国語訳文を作成する．

（1）日本語パターンに当該日本文が適合するか否かを調べる．入力文がパターン内の変数の意味的な制約条件を満足すれば，両者は適合したと判定される．

（2）適合した変数の値（日本語単語）に対して，日中辞典から，パターンで指定された意味属性に適合する訳語（中国語単語）を検索する．日中辞典に適切な訳語が存在すれば，検索は成功する．

（3）上記で得られた訳語を中国語パターンの該当する変数に代入し，中国語訳文を生成する．

### 3.2　評価基準

翻訳文の評価基準は以下の 4 段階とする．

**A)**　文法が正しく，意味が理解できる．
**B)**　文法に不自然なところがあるが，意味が理解できる．
**C)**　文法が間違っているが，意味が大体理解できる．
**D)**　全く意味が理解できない．

### 3.3　評価の例

評価値 A ,B, C, D の例を各々以下に示す．

A 評価の例：

- テスト文 ：
  私の家は駅から近い.
- 使用された結合価パターン対：
  日本語パターン：( "近い")
  　*N*1 "が" *N*2 "に/と/から" 近い
  中国語パターン: ("近")
  　*N*1 "和" *N*2 近
  *N*1 の意味属性：(*すべての意味属性)
  *N*2 の意味属性：(*すべての意味属性)
- 変数の翻訳：家 (*)–>家 ， 駅 (*)–>車站
- 訳文出力：我的家和車站近.

B 評価の例：

- テスト文 ：
  三杯の水を飲む.
- 使用された結合価パターン対：
  日本語パターン：("飲む")
  　*N*1 "が" *N*2 "を" 飲む
  中国語パターン：("喝")
  　N1 喝 N2
  *N*1 の意味属性：(4 人 535 動物)
  *N*2 の意味属性：(746 液体 857 飲物)
- 変数の翻訳：水 (746)–>水
- 訳文出力：喝 3 杯的水.
- B 評価の原因：助動詞がないので，文の全体は不自然である. 普段は「喝了 3 杯的水.」と言う．

C 評価の例：

- テスト文：
  彼の性格が作品に出る.
- 使用された結合価パターン対：
  日本語パターン：("でる")
  　*N*1 "が" *N*2 "に" でる
  中国語パターン：("出現")
  　*N*1 出現 "在" *N*2
  *N*1 の意味属性：
  (* -2671 暦日以外のすべての意味属性 )
  *N*2 の意味属性：(*すべての意味属性)
- 変数値の翻訳：性格 (*) –> 性格
- 訳文出力：他的性格出現在作品.
- C 評価の原因：意味を理解しにくい.
  普通は「他的性格体現在作品.」と言う．

D 評価の例：

- テスト文：
  その子供の親はひどい.
- 使用された結合価パターン対：
  日本語パターン：("ひどい")
  　*N*1 "が" ひどい
  中国語パターン：("出現")
  　*N*1 厉害
  *N*1 の意味属性：(3 主体 2055 出来事 1560 行為 2422 抽象的関係)
- 変数値の翻訳：N1(3):親 –>家长
- 訳文出力：那个孩子的家长厉害.
- D 評価の原因：意味を理解できない.
  普通は「那个孩子的家长無情.」と言う．

### 3.4　評価結果

3.1 節において選択した 200 文に対する，評価結果を表 1 にまとめる．

表 1 では，A 評価が 80 %となっており，作成した日中結合価パターン辞書は，単文の日中翻訳において大変

表 1　日中翻訳の結果

| 評価値 | 結果の割合 |
|---|---|
| A | 80 %(160/200) |
| B | 7.5%(15/200 ) |
| C | 5 %(10/200 ) |
| D | 7.5%(15/200 ) |

有効であることが分かる．すでに述べたように，パターン翻訳において，パターン化の基準は，対象とする言語ペアに依存すると考えられるが，この結果は，日英翻訳のために作成された日本語結合価パターンが，日中翻訳でもかなりの程度使用できることを示している．

これに対して，B 以下の評価となった入力文（20%）は，日英翻訳で定められた日本語結合価パターンのカバー範囲が日中翻訳では適切でないこと，又は，結合価パターン方式の限界を示していることが考えられる．

## 4　考察

### 4.1　日本語結合価パターンのカバー範囲の問題

実験結果に基づき，日本語結合価パターンに適合した日本語文が，対応する中国語結合価パターンでは正しく翻訳できない場合について検討する．

ケース 1：意味的な制約条件のない名詞変数の問題

日英結合価パターンでは，意味的な制約条件の付与されていない変数がかなり存在する．これは，日英翻訳では，特定の格要素の意味属性で英語文型が決定される場合がかなり存在するためである．これに対して，日英翻訳で意味的な制約を不要とされていた変数の中にも，日中翻訳では，意味的な制約条件を付与すべき変数が存在する．これは，日英翻訳と日中翻訳では，訳し分けで重要な要素は同じでないことを示している．

3.3 節の C 評価の例文はこの例に当たる．日本語結合価パターンに対して，3 つの中国語結合価パターンが対応する場合の例を表 2 に示す．

表 2　意味属性により作成できる中国語結合価パターン

| |
|---|
| 1. 出現：*N*1　出現　"在"　*N*2 |
| 2. 体現：*N*1　体現　"在"　*N*2 |
| 3. 出版：*N*1　出版　"在"　*N*2 |

ケース 2：名詞の意味的な制約条件の粒度の問題

変数に対する意味的な制約条件が付与されている日本語結合価パターンでも，その条件が甘く，対応する中国語結合価パターンが複数存在する場合が多くある．試作した結合価パターン辞書では，そのうちの一つしか定義されておらず，誤った訳文が生成される．

3.3 節の D 評価の例文はこの例に当たる．日本語結合価パターンに対して，2 つの中国語結合価パターンが対応する場合の例を表 3 に示す．

表 3　意味属性により作成できる中国語結合価パターン

| |
|---|
| 1. 厉害：*N*1　厉害 |
| 2. 無情：*N*1　無情 |

ケース 1 とケース 2 の問題で翻訳に失敗した入力文は，全体の約 15%(30/200) に相当する．これらの問題を解決するには，現在の日英翻訳用の日本語結合価パターンの変数の意味属性の見直しが必要であること，また，その際，必要に応じて，名詞の意味分類体系をより詳細化する必要がある．

ケース 3：用言訳語選択の問題

日本語結合価パターンのカバー範囲が広く，複数の中国語結合価パターンが対応するような場合でも，一つの汎用的な中国語結合価パターンで済ますことのできる場合がある．例を以下に示す．

日本語結合価パターン：*N*1 が 綺麗

*N*1 の意味属性は：(4 人　468 自然　534 生物　760 人工物　1002 抽象物　2304 自然現象　2564 形状)

この例では，日本語の「綺麗」に対して，中国語の「美丽」,「好看」,「美观」の 3 つの述語が意味的に対応する．従って、これらの意味を訳し分けるには、表 4 のような 3 つのパターンが必要となる。

表 4　意味属性により作成できる中国語結合価パターン

| |
|---|
| 1. 美丽： N1　美丽 |
| 2. 好看： N1　好看 |
| 3. 美观： N1　美观 |

しかし，「好看」,「美观」は，それぞれ特殊な意味での「きれいさ」を表すのに対して,「美丽」は，より一般的な意味での「きれいさ」を表すため，上記の日本語結合価パターンには,「美丽」を使用した中国語結合価パターンを対応させれば，意味の正しい中国語文が作成できる．すなわち，この日本語結合価パターンには，中国語結合価パターン「1. 美丽： *N*1 美丽」を対応させれば，正しい翻訳が可能となる．

実験結果によれば，このようなパターンの改良で正しい翻訳が可能になる入力文は約 3%(6 文/200 文) である．

ケース 4：不適切な名詞訳語の問題

入力文に対して適切な構造の中国語の訳文が作成されているときでも，変数部分（名詞）において不適切な訳語が選ばれていることがある．例を以下に示す．

日本語の名詞「木」は「树」,「木头」,「梆子」などという 3 つの中国語訳語があり，意味的な用法として，それぞれ「樹木」,「材木」,「楽器」という意味属性を持つ．そこで,「木」の意味属性が決まることで，対応する訳語

も決定できる．

ただし，変数化された日本語の単語に対して訳語が絞り込めない場合がある．文献 [4] によれば，このような語が日本語単語の全体の約 5% 存在する．この問題を解決するには，単語の意味属性体系を中国語単語の意味と整合するように見直すことが必要と考えられる．

### 4.2 結合価パターン方式の限界を超える問題

結合価パターンは，述部用言と格要素の意味的な関係を記述する枠組みであり，命題レベルでの，単文の意味を定義する方法として使用される．本節では，このような結合価パターンの限界を超える問題として，副詞的表現の翻訳と時制，相に関連する問題を取り上げる．

#### 4.2.1 副詞の語順の問題

中国語の基本的な語順は「S(主語)+ Adv（副詞) + V(動詞) + Adj(形容詞) +O(目的語)」である．しかし, 動詞, 特有名詞，特殊の強調などにより，語順の例外がある．例えば,「明日学校に行く」を機械翻訳した場合，副詞「明日」は図 2 に示すように四つの場所に置くことができる．

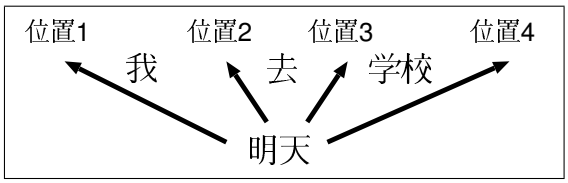


図 2 「明日」の語順

位置 1，位置 2，位置 4 の意味は同じであるが，位置 3 の場合の意味は異なる．このような副詞の語順の問題を解決するには，副詞の要素も含めた文型パターン化が必要である．

#### 4.2.2 時制・相により動詞が選択される問題

日本語では，通常，時制と相は助動詞によって表現される．これに対して, 中国語の動詞は, 動態動詞, 静態動詞，結果動詞に分類され，動詞の種類によってこれらの情報が表される場合がある．特に，動態動詞は助動詞の補佐がないと，文の愛昧さ，および，不自然さが生じる．3.3 節の B 評価の例文はこの問題で正解とならなかったものである．この問題も前項と同様，結合価文法の枠組みを超える問題であり，これを解決するには，助動詞の要素も含めた文型パターン化が必要と考えられる．

## 5 おわりに

本研究では，使用頻度が高い日本語結合価パターン（200 件）を対象に対応する中国語結合価パターンを作成し，日中機械翻訳における結合価パターン翻訳方式の可能性と問題点を検討した．その結果によれば，使用頻度が高く，意味的な訳し分けが必要と見られる日本語単文表現に対して，約 80 %は，正しい中国語訳文が得られることが分かった．従って，日英翻訳用に作成された結合価パターン対の日本語パターンは，日中翻訳のための結合価パターン対の作成においても有効だと言える．

また，翻訳誤りの分析によれば，誤りの大半は，日本語結合価パターンのカバー範囲の不適切さに起因していることが分かった．翻訳誤り 20 %のうちの 15% は，日本語結合価パターンに適合した入力文が，必ずしも対応する中国語結合価パターンで訳すことはできず，意味によってより細かく訳し分けなければならないものであった．この問題を解決するには，中国語の表現構造に着目してそれに対応するように日本語結合価パターン自身を見直すこと，また，適合する日本文の範囲の適正化を図るため，日本語結合価パターン内の変数の意味的な制約条件を見直す必要のあることが分かった．

ところで，副詞の語順の問題や時制，相の問題で正しく訳せないものも 5 %程度存在するが，これらは，結合価文法の枠組みを超える問題であり，これらの問題を解決するには，副詞，助動詞などの表現要素を含むパターン辞書を開発する必要があると思われる．

本研究では，使用頻度の高い日本語結合価パターンの一部を対象に日中結合価パターン辞書を作成したが，日英翻訳用の結合価パターン辞書に収録された日本語結合価パターンは，日中翻訳でもかなり有効であることが分かったので，今後は，残された日本語結合価パターンに対しても中国語結合価パターンを作成し，実験的な改良を行うことにより，日中結合価パターン辞書を実現したい．

## 参考文献

[1] 長尾ほか:自然言語処理,ISBM4-00-010355-5, 岩波書店,1996

[2] 日本語語彙大系 ,NTT コミュニケーション科学研究所, 池原 悟ほか

[3] 西山 七絵ほか：「単文文型パターン辞書の構築」，言語処理学会第 11 回年次大会発表論文集，pp.372-375.2005.

[4] 展瑜ほか：「日中機械翻訳における名詞訳語の選択」，言語処理学会第 9 回年次大会 C4-4 pp.334-337 2003.

[5] 金出地真人ほか：「結合価文法による動詞と名詞の訳語選択能力の評価」，情報処理学会研究報告 2003-NL-153-16 pp.119-124 . 2003-01.

[6] ALT/JE 関連