結合価パターンを用いた日中機械翻訳システムの構築

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

本研究では，IPAL 動詞を対象した日中結合価パターン辞書 (約 5 千件) を使用し，日中機械翻訳システムを試作する．そして，作成した日中翻訳システムを用いて，翻訳テストを行ない，翻訳能力を評価した．翻訳実験結果では，オープンテストにおいて，約 8 割の入力文が正しく翻訳された．また，劣化の評価文を分析し，日中翻訳における日本語結合価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翻訳方式の可能性を明らかにした．

Construction of the J/C Machine Translation System with Valency Patterns

YANG PENG　JIN'ICHI MURAKAMI　MASATO TOKUHISA and SATORU IKEHARA
Faculty of Engineering, Tottori University
{s022061,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

This research is a pilot study on J/C(Japanese/Chinese) mechanical translation system, using a J/C(Japanese/Chinese) valency pattern dictionary (about 5,000 cases) . Based on this experimental translation system ,a translation test was done and the translation ability of this MT system was evaluated . According to the experimental result in the open test, about 80% input sentences were translated in a correct way. With an analysis of incorrect output , several problems of the Japanese valency pattern on the J/C translation were found out ,and the possibility with the valency pattern was clarified .

1　はじめに

近年，日中/中日に関する機械翻訳の研究が盛んである．日中機械翻訳では，統計や用例翻訳方式は既に有効であることが報告されている．しかし，大規模な文型パターン辞書を用いた翻訳方式の検討が少ないため，日中機械翻訳において，効果が不明確である [1]．そこで，本研究では日中機械翻訳の手法の一つである結合価パターン翻訳方式を用いた日中機械翻訳システムを試作する．

具体的には，IPAL 動詞 [2](計算機用日本語基本動詞, 合計 955 語) を対象に日中結合価パターン辞書 (約 5 千件) を作成する [3] [4]．そして，作成した日中結合価パターン辞書を使用し，日中機械翻訳システムを試作する．更に，試作した翻訳システムを用いて，翻訳テストを行ない，翻訳能力を評価する．最後に，翻訳実験に基づき，日中翻訳における日本語結合価パターンの構成法上の問題点と，結合価パターン翻訳方式の可能性を明らかにする．

2　日中結合価パターン辞書

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る．本研究では，既に開発された結合価パターン辞書 (4,903 件)[4] を使用し，システムを構築する．利用する結合価パターン辞書は IPAL 動詞を対象に，日本語結合価パターンを選択する．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し，日中結合価パターン対辞書を構成する．また，使用した「計算機用日本語基本動詞辞書」は基本的な和語動詞 861 語とサ変動詞 94 語を収録している．

2.1　日本語結合価パターン

日中結合価パターン辞書で使用する日本語結合価パターンは日本語語彙大系のパターンである．

日本語語彙大系 [5] 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 6,000 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

属性が掲載されている.

2.2 日中結合価パターン対

日中結合価パターン辞書で使用する中国語パターンは IPAL 動詞と対応する日本語結合価パターンと対応する中国語結合価パターンであり，全部で 4,903 件ある.
日中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" $N2$ "を" 開ける
- 体言の意味属性：
  $N1$：(3 主体)， $N2$：(533 具体物　389 施設)
- 中国語結合価パターン：
  $N1$　打開　$N2$
- パターン作成で参考した日本語例文と中国語訳文
  日本語例文 1：生徒は引出しを開ける.
  中国語訳文 1：学生打開抽屉.
  日本語例文 2：彼は窓を開ける.
  中国語訳文 2：他打開窗户.

2.3 日中単語辞書

結合価パターンを利用するために，体言の意味体系が必要である．本研究では，体言の意味属性を付与した日中単語辞典 (約 1,200 語) を手作業で作成する [6]．例を以下に示す.

日本語の名詞「木」は「树」,「木头」,「梆子」という 3 つの中国語訳語がある．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，対応する訳語が決定できる.

3 翻訳システムの構築

結合価パターンによる翻訳能力を評価するため，図 1 に示すような流れで機械翻訳システムを試作する.

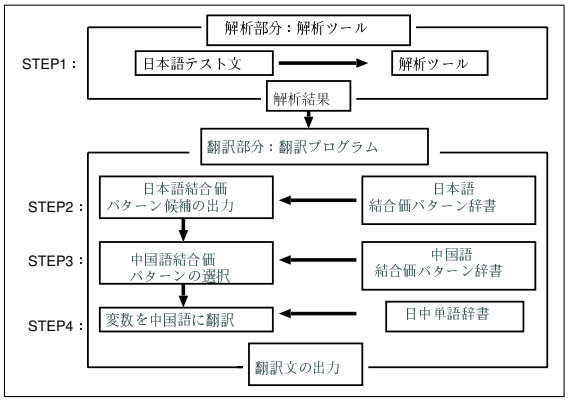

図 1　翻訳実験の流れ

翻訳システムは 2 つの部分で分かれている．解析部分と訳文生成部分である．具体的な翻訳手順と翻訳例を以下に示す.

解析部分

図 1 に示したステップ 1 において，入力された日本語テスト文に対して解析ツールで，形態素解析を行なう.
入力文と解析の例を以下に示す.

- 入力文：中国は日本にパンダを贈る.

形態素解析結果：

1.) 中国 (367 公共機関,385 国家,405 学校,463 領土,30 独立国,98 大学・高専)/は (7530)
2.) 日本 (385 国家,463 領土,30 独立国)/に (7430)
3.) パンダ (537 獣)/を (7430)
4.) 贈る (2386,19 所有的移動)/。([P]0110)

訳文生成部分

上記解析部分で分析した結果を翻訳プログラムに入力する．翻訳プログラムは日中結合価パターン辞書と日中単語辞書を使用し，以下の手順で，各日本語入力文に対する中国語訳文を作成する.

（1）日本語結合価パターンの選択

図 1 に示したステップ 2 では，テスト文と適合する日本語結合価パターンを調べる．同じ用言であり，さらに，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する．また，適合した日本語結合価パターンとパターン ID を出力する．上記に入力したテスト文に対する出力結果を以下に示す.

- 日本語結合価パターン：
  $N1$ が $N2$ を $N3$ に/へ 贈る

$N1$(3 主体) $N2$(533 具体物 1001 抽象物) $N3$(3 主体)

- パターン ID：301429-00

（2）中国語結合価パターンの出力

図 1 に示したステップ 3 では，同じパターン ID の中国語パターンを出力する.

- 中国語結合価パターン：
  $N1$　送给　$N3$　$N2$

（3）単語の翻訳

図 1 に示したステップ 4 では，適合した変数の値（日本語単語）に対して，日中単語辞典から，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 [6].

- 訳語の翻訳：
  中国 (385 国家)→ 中国

日本 (385 国家)→ 日本
パンダ (537 獣)→ 熊猫

（4）訳文の生成

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．訳文の生成結果を以下に示す．

- 中国語訳文：中国送給日本熊猫.

4 翻訳実験

開発した結合価パターン辞書を用いた翻訳システムの翻訳精度を確認するため，オープンテストを行なった．

4.1 オープンテストの入力文

日本語単文集 [7] から，IPAL 動詞で構成する単文を任意に選択し，入力文とする (合計 200 文)．なお，選択した日本文は更に 4 つの条件を満たすように修正する．

(1) 慣用表現を含まない単文.
(2) 文中に副詞句を含まない.
(3) 全ての格要素は結合価パターンに含まれる．
(4) 文末が用言の終止形で終る．

4.2 翻訳実験の評価基準

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) 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.

4.3 オープンテストの例

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 ：
  風は南西から北へ吹く
- 使用された結合価パターン対：
  日本語パターン：(" 吹く")
  $N1$ "が" $N2$ "から, より" $N3$ "に，へ" 吹く
  中国語パターン: ("吹")
  $N1$ "从" $N2$ 吹向 $N3$
  $N1$ の意味属性：(2373 風)
  $N2$ の意味属性：(388 場所)
  $N3$ の意味属性：(388 場所)
- 変数の翻訳：
  風 → 风，南西 → 西南，北 → 北方
- 訳文出力：风从西南吹向北方.

B 評価の例：

- テスト文 ：
  私は叫び声を聞く
- 使用された結合価パターン対：
  日本語パターン：("聞く")
  $N1$ "が" $N2$ "を" $N3$ "から" 聞く
  中国語パターン：("听")
  $N1$ "从" $N3$ 听 $N2$
  $N1$ の意味属性 (3 主体 535 動物)
  $N2$ の意味属性 (1000 抽象)
  $N3$ の意味属性 (3 主体)
- 変数の翻訳：私 → 我，叫び声 → 叫声
- 訳文出力：我听叫声.
- B 評価の原因：結果補語がないので，文の全体は不自然である. 通常「我听到/着叫声」である．

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")
  $N1$ "が" $N2$ "に" でる
  中国語パターン：("出現")
  $N1$ 出現 "在" $N2$
  $N1$ の意味属性：
  (* -2671 暦日以外のすべての意味属性)
  $N2$ 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.
  通常「他的性格体現在作品.」である．

D 評価の例：

- テスト文：
  彼はタバコをやめる
- 使用された結合価パターン対：
  日本語パターン：("やめる")
  $N1$ "が" $N2$ "を" やめる
  中国語パターン：("停止")
  $N1$ 停止 $N2$
  $N1$ の意味属性：(4 人)
  $N2$ の意味属性：(862 たばこ)
- 変数値の翻訳:彼 → 他，タバコ → 烟
- 訳文出力：他停止烟.
- D 評価の原因：意味を理解できない.
  通常「他禁烟.」である．

4.4 オープンテストの結果

4.1 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 128 文 (64%) となっており，作成

表 1　オープンテストの結果

評価値	結果
A	128 文 (64%)
B	32 文 (16%)
C	26 文 (13%)
D	14 文 (7%)

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有効であることが分かった．

これに対して，B 以下 (72) 文の評価となった原因として，日中言語族が違うため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．

5　考察

オープンテストの実験結果に基づき，評価値 B 以下の 72 文の原因を調査する．

5.1　日本語結合価パターンのカバー範囲

本研究で使用した日本語語彙大系の結合価パターンは日英翻訳のために作ったものである．日中機械翻訳において，日本語結合価パターンのカバー範囲の問題を生じる．この問題により，翻訳に失敗した入力文は，全体の 23%(46 文/200 文) である．この問題は 5.1.1 節と 5.1.2 節で解説した 2 つのケースがある．これらの問題を解決するには，現在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り，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と考えている．

5.1.1　意味的な制約条件がない名詞変数 (12.5%)

日本語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数多く存在する．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多く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．評価値 B 以下の 72 文中の 25 文 (全体の 12.5%) はこの問題に起因する．25 文中では，B 評価 9 文，C 評価 8 文，D 評価 8 文だった，また，3.3 節の C 評価の例文もこの例に相当する．

使用したテスト文：彼の性格が作品に出る．
日本語パターン：("でる")
　*N*1 "が" *N*2 "に" でる
*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)
この日本語結合価パターンに対して，表 2 に示す 3 つの中国語結合価パターンが対応する．

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 (現れる)： 　*N*1　出現　"在"　*N*2(-920 出版物)
2. 体現 (性格を表現する)： 　*N*1　体現　"在"　*N*2 (-920 出版物)
3. 出版 (出版する)： 　*N*1　出版　"在"　*N*2 (920 出版物)

　実際に使用した中国語結合価パターン：("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
　中国語で「他的性格出現在作品.」と言わず，意味が理解し難くなるため，C 評価になった．通常「他的性格体現在作品.」である．

5.1.2　名詞の意味的な制約条件の粒度の問題 (10.5%)

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，その条件が広く，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．評価値 B 以下の 72 文中の 21 文 (全体の 10.5%) はこの問題に起因する．21 文中では，B 評価 11 文，C 評価 7 文，D 評価 3 文だった．また，3.3 節の D 評価の例文は，この例に相当する．

使用したテスト文：彼はタバコをやめる．
　日本語パターン：("やめる")
　　*N*1 "が" *N*2 "を" やめる
*N*1 の意味属性：(4 人)
*N*2 の意味属性：(862 たばこ)
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具体的な例を表 3 に示す．
　実際に使した中国語結合価パターン：("停止")
　*N*1 "停止" *N*2
*N*1 の意味属性：(4 人)
*N*2 の意味属性：(862 たばこ)
変数値の翻訳:彼 → 他，タバコ → 烟
「他停止烟.」と翻訳され，意味が理解出来ず，D 評価

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. 停止 (停止する)： *N*1　停止　*N*2(862 たばこ)
2. 禁（禁止する)： *N*1　禁　*N*2(862 たばこ)

になった．通常「他禁烟.」である．

5.2　中国語結合価パターンの用言訳語の選択 (7%)

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にまとめることができる．汎用的な中国語結合価パターンが存在しないため，翻訳に失敗した入力文は 200 文中 14 文 (全体の 7%) だった．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：
(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 に示す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：　*N*1　美丽
2. 好看：　*N*1　好看
3. 美观：　*N*1　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 7%(14 文/200 文) である．

5.3　不適切な名詞訳語

翻訳テストでは，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．

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は文献 [6] によれば，日本語単語の全体の約 15% 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．

5.4　結合価パターン方式の限界

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，命題レベルにおいて，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，時制，相に関連する問題，状態補語の問題と副詞的表現の翻訳問題を取り上げる．

5.4.1　時制・相により動詞が選択される問題 (2.5%)

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．評価値 B 以下の 72 文中の 5 文 (全体の 2.5%) はこの問題に起因する．以下に例を示す．
時制・相の問題で B 評価の例：

- テスト文 ：
  私は叫び声を聞く
- 使用された結合価パターン対：
  日本語パターン：("聞く")
  　*N*1　"が"　*N*2　"を"　*N*3　"から"　聞く
  中国語パターン：("听")
  　*N*1　从　*N*3　听　*N*2
  *N*1 の意味属性 (3 主体 535 動物)
  *N*2 の意味属性 (1000 抽象)
  *N*3 の意味属性 (3 主体)
- 変数の翻訳：私 → 我，叫び声 → 叫声
- 訳文出力：我听叫声.
- B 評価の原因：状態を表す助動詞がないので，文の全体は不自然である. 通常「我听到叫声.」または「我听着叫声.」である．

この問題は，結合価文法の枠組みを超える問題であり，この問題を解決するには，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．

5.4.2　状態補語問題 (3.5%)

日本語では，通常，副詞の状態補語を使わない．例えば，「例 1：島に 虎がいる.」や「例 2：家に テレビがある」などに対して,「島の上に虎がいる」と「家の中にテレビがある」と言わなくても，誤解を生じない．

これに対して，中国語では状態補語がよく使われる．例文に対して，通常,「例 1：在島上 有老虎」と「例 2：在家里 有电视」という。しかし，状態補語は名詞により

変わり，動詞の受身状態も関係があるので，一つの結合価パターンに定義するのは困難である．評価値 B 以下の 72 文中の 7 文 (全体の 3.5%) がこの問題に起因する．

この問題に対して，体言の意味属性をより詳細化分類する必要がある．また，体言間の係り関係や体言と用言の受身関係などの情報により，動詞状態を明確化する必要があると考えている．

5.4.3 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」は以下のように 4 つの場所に置くことができる．

(位置 1)明天 我去学校.
(位置 2) 我 明天 去学校.
(位置 3) 我去 明天 学校.
(位置 4) 我去学校 明天.

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．

6 おわりに

本研究では，IPAL 動詞を対象にした日中結合価パターン辞書を使用し，日中機械翻訳システムを試作した．また，作成した翻訳システムを用いて，翻訳実験を行ない，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．

オープンテストによれば，200 文の入力文に対して，128 文は正しい中国語訳文が得られることが分かった．その結果から，開発した日中結合価パターン辞書は日中機械翻訳システムの構築に効果があることが分かった．また，従来の研究 [3]，[4] と合わせ，結合価パターン翻訳方式は単文に対して日中機械翻訳に有効だと言える．

また，翻訳誤りの分析結果によって，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 72 文のうちの 46 文 (全体の 23%) 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ことが必要であると考えられる．また，助動詞の問題や結果補語や時制，相の問題により正しく訳せないものも 26 文 (全体の 13%) 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，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．今後は，改良を行うことにより，より精度が高い日中機械翻訳システムを実現したい．

また，体言の意味属性を付与した日中辞典が存在しないため，本研究では，手作業で意味属性を考えた日中辞典を作成した．しかし，登録した単語の数が少ないため，より規模が大きい翻訳テストを実現できなかった．今後は，意味属性を付与した日中辞典を電子化したいと考えている．

参考文献

[1] 長尾 真ほか:自然言語処理, 岩波書店,1996

[2] 情報処理振興事業協会. 計算機用日本語基本動詞辞書, 1999

[3] 楊 鵬ほか：「結合価パターンを用いた日中機械翻訳方式の検討」, 言語処理学会第 12 回年次大会発表論文集，pp.264-267.2006.

[4] 楊 鵬ほか：「日中機械翻訳対する結合価パターン翻訳方式の応用」, 言語処理学会第 13 回年次大会発表論文集，pp.79-82.2007.

[5] 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

[6] 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.

[7] 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.

[8] 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.