日中機械翻訳に対する結合価パターン翻訳方式の応用

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，言語表現の意味を等価的に変換する方式として，パターンを用いた翻訳方式の研究が行われている [1] [2]．日中機械翻訳では，既に語義数の多い一部の用言の結合価パターン辞書 (300 件) を試作し，単文の翻訳において，結合価パターンの翻訳方式が有効であることが報告されている [3]．しかし，結合価パターンによる方式は，基本的な用言の訳し分け効果がまだ不明確である．更に，小規模の結合価パターン辞書しか開発しなかったため，パターン翻訳に対する信頼性が低い．

そこで，本研究では，IPAL 動詞 (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．さらに，作成した日中結合価パターン辞書を用いて，翻訳能力を評価する．最後に，オープンテストを行い，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　日中結合価パターンの試作

2.1　結合価パターン

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．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において，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．

本研究では，既に開発された日本語語彙大系を参考し，IPAL 動詞と対応する日本語結合価パターンを対象に，対応する中国語結合価パターンを作成する．

2.2　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 [4] 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．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：
  優勝する　(状態 受身不可)
  *N*1　"が"　*N*2　"に/で"　優勝する
  *N*1(3 主体 535 動物) *N*2(1001 抽象物 1236 人間活動)

2.3　IPAL 動詞

IPAL 辞書 [5] は計算機による日本語処理のために作成された辞書である．「計算機用日本語基本動詞辞書」，「計算機用日本語基本形容詞辞書」，及び，「計算機用日本語基本名詞辞書」から構成される．

本研究で，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．

2.4　中国語結合価パターンの作成方法

以下に示す 4 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
   日本語語彙大系 [4] の結合価パターン辞書から, IPAL 動詞に対応する日本語結合価パターン（4,903 件）を選択する．
2. 日本文の選択
   上記で選択された各日本語結合価パターンに対して，日本語単文集 [6] から，適合する日本文を 2 文選択し，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) 副詞句を含まない．
   (2) 形容詞句と名詞句は対象内とする．
   (3) 全ての格要素は結合価パターンに含まれる．
   (4) 用言の終止形で終る．
3. 中国語訳文の付与
   選択した日本文に対応する中国語訳文を付与する．
4. 中国語結合価パターンを作成
   上記 2. と 3. 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 (4,903 件) 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：
  *N*1　"が"　*N*2　"を"　開ける

- 参考した日本語例文と中国語訳文
  日本語例文 1：生徒は引出しをあける.
  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.
- 中国語結合価パターン：
  *N*1　打開　*N*2
- 体言の意味属性：
  *N*1：(3 主体),　*N*2：(533 具体物　389 施設)

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

結合価パターンによる翻訳能力を評価するため，オープンテストを行なった．実験の流れを図 1 に示す.

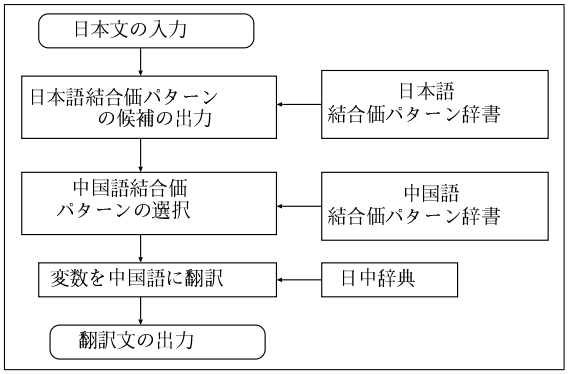

図 1　翻訳実験の流れ

具体的な翻訳手順を以下に示す.

1. 入力文の選択：

日本語単文集 [6] 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 100 文). 但し，パターンの作成で参照した日本文と異なる単文とする．更に，2.4 節の「日本文の選択」と同じ手順で，日本文を修正する.

2. 中国語訳文の作成：

日中結合価パターン辞書を使用し，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）入力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.

（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 [7].

（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 ：
  私はラケットでボールを打つ
- 使用された結合価パターン対：
  日本語パターン：("打つ")
  　*N*1 "が" *N*2 "を" *N*3 "で" 打つ
  中国語パターン: ("打")
  　*N*1 "用" *N*3 打 *N*2
  *N*1 の意味属性：(4 人)
  *N*2 の意味属性：(552 動物 (部分) 921 遊び道具・運動具　2020 打ち 533 具体物)
  *N*3 の意味属性：(-955 棒　863 建造物)
- 変数の翻訳：
  　私 → 我，ボール → 球，ラケット → 球拍
- 訳文出力：我用球拍打球.

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：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")
  　*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.
  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")
  *N*1 "が" *N*2 "を" やめる
  中国語パターン：("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
- 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.

3.4 評価結果

3.1 節において選択した 100 文に対する，評価結果を表 1 にまとめる.

表 1 日中翻訳の結果

評価値	結果
A	86 文
B	3 文
C	10 文
D	1 文

表 1 では，A 評価が 86 文となっており，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有効であることが分かる.

これに対して，B 以下の評価となった原因（14 文）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.

4 考察

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，評価値 B 以下の 14 文について検討する.

4.1 日本語結合価パターンのカバー範囲の問題

ケース 1：意味的な制約条件が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．評価値 B 以下の 14 文中の 5 文はこの問題に起因する.

また，3.3 節の C 評価の例文もこの例に相当する．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具体的な例を表 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 (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)

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 14 文中の 4 文はこの問題に起因する.

また，3.3 節の D 評価の例文はこの例に相当する．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具体的な例を表 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. 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)

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 9 文 (9/100) に相当する．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.

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.

4.2.1 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．評価値 B 以下の 14 文中の 3 文はこの問題に起因する．また，3.3 節の B 評価の例文はこの問題に相当する．

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．

4.2.2 状態補語問題

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，普段では，「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．この問題により，翻訳に評価値 B 以下の 14 文中の 2 文に相当する．

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である．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．

4.2.3 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」は以下のように四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．

5 おわりに

本研究では，日本語基本動詞の日本語結合価パターン（4,903 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，基本動詞の結合価パターンでは，日本語単文表現の 100 文に対して，86 文は正しい中国語訳文が得られることが分かった．また，先行研究 [3] により，使用頻度が高い結合価パターン対に日中機械翻訳に対して有効であるので，結合価パターン翻訳方式は全体的に日中機械翻訳にと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 14 文のうちの 9 文は，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，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 5 文が存在するが，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．

本研究では，基本動詞の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た．実験により，結合価パターン翻訳方式は，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，改良を行うことにより，より精度が高い日中結合価パターン辞書を実現したい．

参考文献

[1] 長尾 真ほか:自然言語処理, 岩波書店,1996

[2] 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.

[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] 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

[5] 情報処理振興事業協会. 計算機用日本語基本動詞辞書, 1999.

[6] 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.

[7] 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.