062023 結合価パターンを用いた日中機械翻訳システムの構築

計算機工学講座　C　楊鵬

1　はじめに

高品質な機械翻訳を目指し，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なわれている．日英翻訳では，既に網羅的なパターン辞書が開発され，機械翻訳において，パターン翻訳方式は有効であることを報告されている．しかし，日中翻訳において，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，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] 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．

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 [5] のパターンである．日本語語彙大系は，「構文体系」と「意味体系」から構成される．

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　$N2$

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 [7] を試作する．

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．

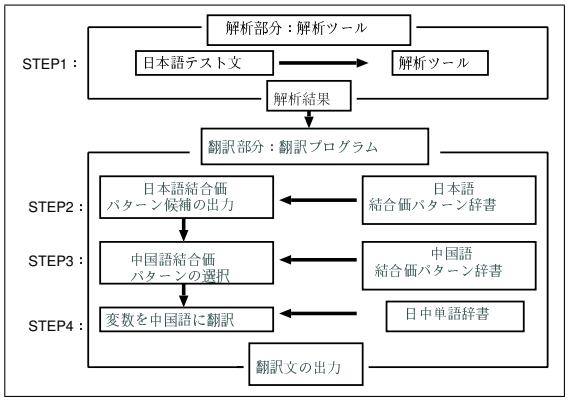

図 1 翻訳実験の流れ

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．

（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．

（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].

（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．

4　翻訳実験

4.1　オープンテストの入力文

日本語単文集 [8] 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 の例を以下に示す.

- 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.

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 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%)

表 1 では，A 評価が 128 文 (64%) 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った．

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

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．

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，特定の格要素の意味属性で日本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，日本語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，日中両言語では，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．例を以下に示す．
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")
　$N1$ "が" $N2$ "に" でる
意味属性:$N1$(主体)，$N2$(創作物，目録，通信械器)
　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. 体現 (性格を表現する)： 　$N1$ 体現 "在" $N2$
2. 出版 (出版する)： 　$N1$ 出版 "在" $N2$
3. 出現 (現れる)： 　$N1$ 出現 "在" $N2$

　実際に使用した中国語結合価パターン：("出現")
　$N1$ 出現 "在" $N2$
変数値の翻訳：性格 → 性格，作品 → 作品
「他的性格出現在作品.」と翻訳され，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．例を以下に示す．
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")
　　$N1$ "が" $N2$ "を" やめる
$N1$ の意味属性：(4 人)
$N2$ 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．

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)： 　$N1$ 停止 $N2$(862 たばこ)
2. 禁（禁止する)： 　$N1$ 禁 $N2$(862 たばこ)

　実際に使した中国語結合価パターン：("停止")
　$N1$ "停止" $N2$
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 評価になった．通常「他禁烟.」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，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，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] 池原 悟ほか :「日本語語彙大系」
[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 pp.334-337，2003.
[7] 楊 鵬ほか :「結合価パターンを用いた日中機械翻訳システムの構築」情報処理学会研究報告会, pp.121-126, 2008.
[8] 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，pp.372-375，2005.