入力文に対する結合価パターン対の選択方法について

吉田 真司　池原 悟　村上 仁一
鳥取大学工学部知能情報工学科
{yosita,ikehara,murakami}@ike.tottori-u.ac.jp

1 はじめに

機械翻訳では複数の訳語を持つ語の訳語選択が問題となる。解決の手法として結合価パターンは訳語選択の精度向上に有効である。しかし、実際の文は語順変更、省略等により現状の登録パターンで結合価パターンを適用できない事がある。特に省略の場合、誤ったパターンを導き出す場合が多数あり、難しい問題となる。よって候補パターンを出来る限り絞り、正しい結合価パターンを導き出す事が、結合価パターンを用いる機械翻訳で重要となる。対処法として、文献 [2] の辞書登録パターンから可能な展開形を複数派生させ、展開形と入力文を照合する方法が提案されている。しかし、省略には対応できない。

本研究は、語順変更、省略に対応した精度の高い結合価パターン対の選択方法を考案する。入力文の形態素解析を行い、格要素ごとに分割した文を、結合価パターンと照合する事を基本とする。さらに格要素の名詞の深さ、格助詞の格の種類に着目した点数付けによって省略に対応する事を可能とする。

2 結合価パターンによる意味的制約

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た物で、用言と格要素間に意味的な制約をつける事で訳語選択の精度を高める。本研究では、日本語語彙大系 [3] の「構文意味辞書」の結合価パターンを基本にする。結合価パターンは一般表現と慣用表現を合わせ、約 16,000 件の日本語文型パターンにまとめている。結合価パターンの例を表 1 に示す。

表 1: 結合価パターンの例（一般表現と慣用表現）

	結合価パターン	対応する英文
一般	$N1$(性質) が大きい	$N1$ be remarkable
	$N1$(主体) が $N2$(人) を絞る	$N1$ grill $N2$
慣用	$N1$ が名を売る	$N1$ make $N1$-self famous
	$N1$(人) が知恵を絞る	$N1$ rack $N1's$ brains

一般表現は用言と一つ以上の格要素で規定する。慣用表現は、名詞が意味属性でなく直接単語の字面で規定され、主に特殊な表現で使う。

2.2 一般名詞意味属性体系

本研究では、結合価パターンの名詞の意味属性を調べるために、一般名詞意味属性体系 (図 1) を用いる。一般名詞意味属性とは、単語の意味により体系的に分類、整理した語彙集である。日本語語彙大系に登録してある結合価パターンにある名詞は、一般名詞意味属性体系によって意味属性を付けられている。

本研究で使用する一般名詞意味属性体系は、約 40 万語の名詞を、最大 12 段の木構造を構成する 2,710 の意味属性に分類している。また、一般名詞意味属性体系は、木構造を基本構成とし、上位の属性を持つ名詞は、木構造において自分より下位の名詞の意味属性を包含する性質がある。上位の意味属性が下位の意味属性を包含する事と名詞の持つ意味属性を利用し、結合価パターンは動詞の持つ意味を規定する事ができる。

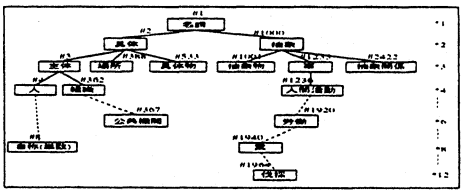

図 1: 一般名詞意味属性体系の一部

2.3 単文の意味的整合性の判定

文の意味的整合性を判定する方法を具体例で示す。判定は結合価パターンと、名詞意味属性体系の上位の意味属性が下位の意味属性を包含する事を利用する。

以下の結合価パターンの例について考える。

$N1$(主体) が $N2$(人) を絞る。

上記の例に対して次の 2 文を考える。

例文 1) (私) が (人員) を絞る。
例文 2) (彼) が (頭) を絞る。

$N1$ にあてはまる名詞は、例文 1 においては (8 : 自称 (単数))、例文 2 においては (22 : 他称 (単数)) であり、共に (4 人) 以下の属性なので正しい答となる。しかし $N2$ にあてはまる名詞は、例文 1 においては (4 人) 以

下の属性の名詞 (119 ：成員) なので正答になるのに対し、例文 2 では (554 頭) という人以下の属性の名詞ではないので別の意味のパターンであると判定される。

3 結合価パターンの照合

3.1 格助詞の点数付け

本研究では、複数出現する結合価パターンの候補を絞るために格要素に着目し、名詞と付随する格助詞を利用する事で省略に対応する照合方法を提案する。

省略のある文は、省略部分によって重要性が変わってくる。例えば「が」格なら省略してもさほど文全体の意味に影響しない事が多い。しかし、「を」格は目的格となる場合が多く、省略すると文全体の意味が曖昧になり、最悪の場合意味不明の文となる。

よって、格助詞の種類によって、重要性が違うと仮定し、格助詞に点数付けを行う。また、格助詞が付随する名詞は、意味属性の深さが深ければ深い程名詞の意味を制約する働きがあると考えられる。意味を制約する事で候補パターンを絞り込む事ができる事から名詞の深さを点数として用い、格助詞との積で点数を決定する。候補パターンの中で格要素の点数合計が最高得点のパターンを最適なパターンとする事で候補文を一意に絞る。

具体的には以下の計算式を用いる。

$D = \Sigma\ (A * B)$ … (1)

D : パターンの点数

A : 名詞の属性の深さ (1〜12) B : 格助詞の点数

表の格助詞の点数付けについては、[3] にある結合価パターンでの格助詞の使用頻度によって点数を規定する。例えば「が」格は頻繁に使用される割に省略しても文章全体の意味に影響する事は少ないので点数は低くする。基本的に使用頻度の低い物を重要度が高いとし、点数付けを行う。また、各パターンにおいての省略部分については減点を行う。

3.2 格助詞の配点について

格助詞の使用頻度調査は [3] を利用する。[3] に登録してあるパターン数が 10 個の自動詞と他動詞から、各 5 種類を選び、用いられている格助詞の使用頻度を調査する。格助詞の使用頻度を元に算出した格助詞の配点を表 2 に示す。「が」「に」「を」格以外は使用頻度が低い事と、使用頻度の高い格助詞でも重要度が高い場合がある事を考慮し、点数の調整は点数の開きが少ないように行った。減点については、この加点での最低点数が減点の最高点数と同じになると、減点による弊害が発生するため、調整をした。また、減点によって誤った結合価パターンを出来る限り導き出さないように配点を行っている。

表 2: それぞれの格助詞の配点

	が格	に格, を格	で格	へ格	の格	その他
加点	3 点	3.5 点	4 点	4 点	4.5 点	5 点
減点	0.5 点	0.5 点	0.8 点	1 点	1.2 点	1.5 点

3.3 点数計算の例

点数計算の例を示す。

例 1 : 勉強会を開く。

候補 1 : (主体) が (全て) を開く。

候補 2 : (主体) が (式、行事) を開く。(正解)

1 の計算 (深さ 1*3.5 点)-(深さ 3*0.5 点)=2 点

2 の計算 (深さ 6*3.5 点)-(深さ 3*0.5 点)=19.5 点

どちらの候補パターンも、「が」格が省略されているため、「が」格の分だけ減点となる。加点は「を」格の分で加点、減点が 1 つずつとなる。点数計算では 2 のパターンが正解となる。1 のパターンは「〜を開ける」という意味に対し、2 のパターンは「会を開く」という意味に取れる事からも 2 が正解となる。

次に、減点の有効性を示す例を述べる。

例 2 : 大会開催が決まった。

候補 1 : (全て) が決まる。

候補 2 : (全て) が (全て)[から、より、で] 決まる。

1 の計算 (深さ 1*3 点) ＝ 3 点

2 の計算 (深さ 1*3 点)-(深さ 1*1.5 点) ＝ 1.5 点

減点は、余分な格要素が入っている場合を除外できるため、有効である。ただし、省略を考慮する場合、パターン 2 もありうる。本研究では、出来る限り省略が無く、合致するパターンを正答とする。

4 受身、使役文への対処方法

結合価パターンは能動態で登録されている。受身や使役の文を結合価パターンに対応させるためには、単純に受身、使役文を結合価パターンに登録する方法や既存の結合価パターン自体を変形して入力文に対応させる方法等がある。本研究では入力文の格助詞を変更して平叙文で認識する。一時的に文章を平叙文に直す事により、入力文は結合価パターンに対応し、受身、使役文に対しても結合価パターンを用いる事ができる。

格変化対応表を表 3 に示す。対応表は受身、使役文を平叙文に直す際に最もよく使う格変化を基本とする。ただし、「に」格についてはかかる名詞が主体 (注釈あり) となる場合は格変化せず、「を」格についてはかかる名詞が主体となる場合のみ「が」格に格変化する特別規則を設けている。

例 : 契約が交わされている。

→契約を交わす。(「が」格→「を」格に変化)

表 3: 受身、使役の格変化対応表

受身、使役	が格	に格	を格	と格	から格
平叙文	を格	が格	を格	と格	が格

注：一部例外あり。具体的な適応範囲は (4 人) 以下と (534 生物) 以下の意味属性の名詞。

5 評価実験の方法

語順変更、省略、受身、使役文に対応した点数計算の結合価パターン対応精度を調べる。方法として、新聞記事データより文章を抽出し、点数計算する事で適切な結合価パターンを導き出す精度を実験的に調べる。

5.1 実験の手順

1. 毎日新聞 95 年度記事から入力文を上から 300 文抽出し、単文に加工する。
2. 単文を形態素解析する。
3. 形態素解析済みの文を今回の法則に従って作成したプログラムにかける。
4. プログラムで導き出した結合価パターンが正しいかを人手で判断する。

また、本研究では以下のルールを加える。

a：副詞、感動詞は結合価パターンと無関係なので除外する。

b：可能、尊敬については考慮しない。ただし、可能は受身と意味が似通っている場合、受身として判別する。

c：「は」「も」格は、基本的に「が」格扱いとする。

5.2 人手による評価基準

評価は入力文と結合価パターンの意味を見て比べ、意味が概ね同じ物であったなら正答とする。本研究では、正答であったとしても、複数の場合がある事が認められたので、いくつかの場合に分けて評価する。

1) ◎：導き出したパターンの意味が正しい場合
2) ○：導き出したパターンは正しいが、類似のパターンが他にも存在する場合
3) △：導き出したパターンの意味で正しいが、別の意味のパターンでも意味が通る場合
4) ＊：形態素解析ミスによる意味の取り違いの場合か、格助詞が複数の意味を持つ為に誤ったパターンを導いた場合
5) ×：点数計算が不適切だった場合

正答率は△までを正答とし、＊と×は誤りとする。

6 実験の結果

6.1 格変化、点数計算の有効性

まず、正答率についての結果を表 4 に載せる。本研究の成果を明確にするために、比較材料として、3 種類のデータを用意する。

a：格変化、点数計算無し：名詞の意味属性と付随する格助詞、動詞の照合のみでパターンを絞る場合。複数の候補が残った場合は正解が候補内に一つあるとして、正解を選ぶ確率を加える。

b：点数計算のみ：受身に対しての処置を行わず、複数の候補パターンが出た場合は点数計算によって候補パターンを一意に絞る場合。

c：格変化、計算使用：受身に対して格変化を行い、複数の候補パターンが出た場合は点数計算によって候補パターンを一意に絞る場合。

ただし、この正答率は、全てパターン無しと判定された文は除き、慣用表現を用いている文についても除く。パターン無しの文は 58 文、慣用表現の文は 8 文なので、正答率の判定に用いた文は 234 文である。

表 4: 正答率比較 1

格変化、点数計算無し	点数計算のみ	格変化、計算使用
81.9 % ((191.54)/234)	86.8 % (203/234)	94.0 % (220/234)

6.2 格助詞、意味属性の有効性

格助詞と名詞属性の両方を用いる事の有効性を示すために、参考として以下の 3 通りの計算の比較を行う。

a：格助詞の点数のみでの計算
b：名詞の意味属性の深さのみでの計算
c：意味属性の深さと格助詞の点数の積の計算

結果を表 5 に示す。配点については前述した点数をそのまま用い、受身、使役に対応した格変化も用いる。ここで複数候補の正答率とは、複数の候補パターンが残った時に点数計算によって正解を導き出した正解率を表す。ただし、点数計算の成功率だけを示す。

表 5: 正答率比較 2

	格助詞のみ	属性深さのみ	併用
複数候補の正答率	68.9 % (42/61)	93.4 % (57/61)	96.7 % (59/61)
全体の正答率	86.8 % (203/234)	93.2 % (218/234)	94.0 % (220/234)

6.3 正答と誤りの詳細

格助詞のみを用いた場合は、格助詞単独用に点数の調整をしていないため、精度はそれほど高くはない。だが、意味属性の深さを用いる場合は、両方を用いた場合に近い結果となる。結果より格助詞、意味属性はパターン選択精度向上に有効である事が分かる。次に、全ての正答と誤りの詳細なデータを表 6 に示す。

表 6: 格変化、計算を使用した場合の詳細

◎	○	△	＊	×
85.0 % (199/234)	5.6 % (13/234)	3.4 % (8/234)	5.1 % (12/234)	0.9 % (2/234)

完全に正答と言えるのは◎だが、○や△も正しい答えではあるので総合的には正答としている。

a) ◎の例：勉強会を開く。
候補 1：(主体) が (全て) を開く。
候補 2：(主体) が (式、行事) を開く。(正解)
1 の計算 (深さ 1*3.5 点)-(深さ 3*0.5 点)=2 点
2 の計算 (深さ 6*3.5 点)-(深さ 3*0.5 点)=19.5 点
点数計算上では候補 2 の方が候補 1 より高いので候補 2 を正しいパターンとする。候補 1 は (箱、扉等の開ける物) を開くという意味に対し候補 2 は (会議、会合等) を開く (開催する) という意味なので意味的に入力文と合致するのは候補 2 となる。よって点数計算で候補 2 を選出するのは正しいので、◎の評価となる。
b) ○の例：大会開催が決まった。
候補 1：(全て) が決まる。
候補 2：(全て) が (全て)[から、より、で] 決まる。
3.2 章で述べたパターンで、類似したパターンがある場合にあたる。1、2 とも、どちらの意味とも取れる。
c) △の例：軍隊は全滅する
候補 1：(全て) が全滅する (3 点)
候補 2：(主体 動物) が (全て) を全滅する (8.5 点)
この文の場合、候補 2 を正解としており、確かに候補 2 の意味で取る事は可能である。しかし、本来は別の意味での候補 1 である可能性もある。○とは異なり、明らかに意味が違うが、両方の意味で取れる場合のパターンである。
＊、×については考察で述べる。

7 考察

今回の実験で、正答率は高い数値を得る事が出来た。しかし、一部誤ったパターンを導いた場合があったので表 7 に示す。7.1 は、パターン未検出も含む。

表 7: 誤ったパターンを導き出した場合

	原因	文数
7.1	「は」「も」格が別の性質を持つ場合	3.8 % (9/234)
7.2	点数計算が不適切だった場合	0.9 % (2/234)
7.3	格の変化に対応しきれなかった場合	0.9 % (2/234)
7.4	形態素解析のミスがあった場合	0.4 % (1/234)

7.1〜7.4 の失敗パターンを示す。前述した＊の例は 7.1,7.3,7.4 に対応し、×は 7.2 に対応する。

7.1 「は」「も」格が別の性質を持つ場合

例：首相訪米は予定通り行う。

「は」格が「を」格の性質を持つため、「が」格として認識すると、「首相訪米」は主格でないので誤りとなる。対処としては「は」「も」格に対して新たな法則を見出す必要がある。該当する全ての格助詞の性質を持たせる事を試みようとしたが、「が」格でしかあてはまらない場所に「を」格が当てはまる等、問題が多い。

7.2 点数計算が不適切だった場合

例：産廃処理工場に男の声の電話を二回受けた。

候補 1：$N1$(主体) が $N2$(全て) を $N3$(主体)[から、より]$N4$(衣料 容器) で受ける。(正解に近い)
候補 2：$N1$(創作物 芸 出版等 興行等) が $N2$(主体) に受ける。
候補 1 の計算：(深さ 1*3.5 点)-(深さ 3*0.5 点)-(深さ 3*1.5 点)-(深さ 8*0.8 点)=-8.9 点
候補 2 の計算：(深さ 3*3.5 点)-(深さ 11*0.5 点)=5 点
意味上では候補 1 が正解となるが、省略が多いとみなされ、大幅に減点となるため候補 2 が選出される。今回の実験の場合、格要素が多いパターンでは、減点が多くなる欠点が残る。点数を調整すれば改善は可能であるが、今回のパターンにしか対応できず、逆に対応不可パターンが現れる弊害がある。電話を受けるという意味では候補 1 の方が近いと考えられるが、実際はさらに適したパターンを登録すべきだと考えている。

7.3 格の変化に対応しきれなかった場合

例：花を咲かせる。

使役文ではあるが、「花を咲かせる」から「花が咲く」に変わらなければならず、前述した法則では例外にあたる。よって、誤りとなる。

7.4 形態素解析のミスがあった場合

例：規模が均質に近い。

「均質に」の部分が「均質」と「に」というように分けて取られず、「均質に」という塊で取られてしまうため、判定が出来ない。

8 結論

実験結果より、格の深さに注目した点数計算は、省略において出現した複数の結合価パターン候補を絞る事に有効である事が分かり、精度の高い照合を可能としたと言える。しかし、複数の性質を持つ格や受身、使役での格変化の対応について改良の必要がある。また、特殊な文にも別法則を用いる必要がある。例として「は」格を用いる二重主格文への対応やさまざまな格に変化する「も」格への対処を行えばさらに 2、3 %程度の精度向上が可能だと思われる。また、単文に対しては高い精度で正しい結合価パターンを得る事が出来たが、重文、複文については対応が取れない可能性があるので、今後、検討して行きたい。

参考文献

[1] 金出地真人：「結合価パターン辞書を用いた訳語選択の精度」, 鳥取大学工学部知能情報工学科卒業論文 (2001)
[2] 白井, ほか 4 名：「入力文と結合価パターン対辞書の照合に関する一手法」, 言語処理学会第 5 回年次大会発表論文集, pp.80-83 (1999)
[3] 池原, ほか 6 名：日本語語彙大系, 岩波書店 (1997)