IPAL動詞辞書を用いた日本語語順の推定

徳永健伸, 田中穂積

東京工業大学工学部情報工学科

概要

本稿では, 動詞の結合価情報を利用して日本語単文の語順を推定するモデルについて述べる. 日本語の単文の語順について, 文末の動詞を限定する力が強い後置詞句が先行するという言語学者の指摘がある. 我々は, 後置詞句が動詞を限定する力を動詞の結合価情報から計算し, これに基づいて語順を推定するモデルを提案する. 具体的には, 意味素性と格助詞の特定の組合せを必須格としてとる動詞の数で, この限定力を定量化する. 実際にIPAL基本動詞辞書の結合価情報を使用してモデルの妥当性を検討した.

1 はじめに

日本語は語順が比較的自由な言語であるといわれているが, 語順にまったく制限がないわけではない. 特に述部の語順に関しては言語学者による多くの研究があり, かなり厳格な語順規則があることが明らかにされている [6]. 一方, 述部にかかる要素の語順については述部の語順ほど厳格な規則は見い出されていない.

佐伯は小説67ページ中の文を手作業で分析し, 補語, すなわち, 名詞句と格助詞の組の語順について以下の9つの傾向を抽出している [2].

(1) 位格(ニ, デ, カラ, ヲ)は他の格に先行する

(2) トキの位格はトコロの位格に先行する

(3) ガは位格を除く他の格に先行する

(4) 与格のニは対格のヲに先行する

(5) カラは着格のニ, へに先行する

(6) 長い補語は短い補語に先行する

(7) 文脈指示を含む補語は先行する

(8) 補語同士がかかり受けを構成する場合, かかりの補語が先行する

(9) 慣用句では, 特定の補語は動詞の直前に位置する

佐伯はこれらの傾向を成分条件に基づく語順傾向と構文条件に基づく語順傾向の大きく2つに分類している.

成分条件に基づく語順傾向とは, 補語自身の意味, 機能に備わった支配条件に基づいて生じる語順傾向のことで, (1) から (5) がこれにあたる. これらの傾向をまとめると, おおよそ次のようになる.

位格(トキ > トコロ) > 主格 > 与格 > 対格

ただし, X > YはXがYに先行することを意味する.

一方, 構文条件に基づく語順傾向とは, 補語の構文的な意味, 機能に備わった支配条件に基づいて生じる語順傾向のことで, (6) から (9) がこれにあたる.

佐伯は (1) から (5) の成分条件に基づく語順傾向については, かかりの深さと広さという観点から, (6) から (9) の構文条件に基づく語順傾向については, かかり先のあいまい性という観点から語順傾向の必然性を説明している. 次章では語順傾向に影響を与える佐伯のかかりの広さと深さについて説明し, かかりの広さに基づく語順の推定モデルを提案する.

2 語順の推定モデル

2.1 かかりの深さと広さ

日本語の単文は単純化すると後置詞句並びの後に述語が続くと考えることができる. 本稿では後置詞句の助詞としてガ, ヲ, ニ, カラ, へ, ト, ヨリ, デの8つの格助詞, 述語としては動詞のみを考える. また, 時制, アスペクトは扱わない. すなわち, 我々が扱うのは格助詞によって有標化された名詞句の並びの後に動詞が来るような単文である.

我々の関心は, このように単純化した日本語の単文について, かかり要素 (後置詞句) がどのような順序

に並ぶかという問題である．佐伯はこの問題をかかりの深さと広さという概念を用いて説明している [2]．

かかりの深さとは，直観的にはかかり要素と述部の距離を表し，より文末の受けにかかる方がかかりの深さが深いという．かかりの深い要素のほうが先行しやすいというのが佐伯の観察である．これは非交差の原則とも関係がある．また，日本語の述部の語順にはかなり厳格な規則があるので [6]．述部の要素を語順によって階層化し，かかり要素が述部のどの階層にかかるかによって，かかりの深さの絶対的な指標を求めることができる．かかりの深さという概念は述部の階層構造を前提としているので，述部に動詞のみを仮定する我々の考察対象では役にたたない．

一方，かかりの広さとは，かかり要素がどの程度述部を限定するかという概念であり，述部を厳しく限定するほどかかりの広さが狭いと考える．たとえば，「東京カラ」や「京都へ」のようなかかり要素は，何らかの移動を表す動詞が文末に来ることを予想させるが，「私ガ」のようなかかり要素は，動詞の限定がよりゆるやかである．かかりの広さは述部に動詞しか含まない単文にも適用できるため，以下ではかかりの広さを考察の対象とする．

佐伯は動詞が文末に位置するという日本語の特徴をふまえ，かかりの深さが深く，かかりの広さが広いかかり要素から順に提示する方が聞き手にとって望ましいとしている．いいかえれば，より制約の少ないかかり要素を先に提示することになる．我々の目的は佐伯のかかりの広さを定量化し，それを用いて語順の推定モデルを構築することである．次節では，動詞の結合価情報に基づくモデルについて述べる．

2.2 結合価行列と優先度行列

日本語の動詞は，動詞によって必須要素としてとる格助詞の種類，および格助詞と結び付く名詞句の意味的な性質が異なっている．動詞がどのような意味的性質をもつ名詞句をどの格にとるかを規定するものを，その動詞の結合価パタンと呼ぶ．また，名詞句の意味的性質を表現するために，あらかじめ用意したいくつかの意味素性を用いて表す．

我々のモデルでは，意味素性と格の各組合せについて，その組合せを結合価パタンに含む動詞の数でかかりの広さを定量化する．意味素性と格の組合せは後置詞句の特徴を規定すると考えることができるので，このモデルによって後置詞句のもつかかりの広さの定量的な値を計算することができる．したがって，どの後置詞句が先行しやすいかという指標を与えることができる．以下では，このモデルを定式化する．

ある動詞がある格をとらないときは，その格が仮想的な意味素性 NON をとると考えると，結合価パタンは N_p 項行ベクタで表現できる．これを結合価ベクタと呼ぶ．ただし，N_p は格の数で，ここでは $N_p = 8$ である．結合価ベクタにおける要素の位置は考察の対象としている 8 つの格助詞に対応し，ベクタの要素は意味素性となる．一般に 1 つの動詞は統語的，意味的な違いによって複数の結合価ベクタをとることがある．以下，「動詞」は結合価パタンの違いを考慮したものであるとする．したがって，各動詞は唯一の結合価ベクタを持つことになる．

ここで，N_v 個の動詞について，各動詞の結合価ベクタを縦方向に連接した N_v 行 N_p 列の行列 VM を考える．これを結合価行列と呼ぶ．

$$VM = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1N_p} \\ s_{21} & s_{22} & \cdots & s_{2N_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N_v1} & s_{N_v2} & \cdots & s_{N_vN_p} \end{bmatrix}$$

次に，各意味素性について，その意味素性 N_v 個を要素とする N_v 項行ベクタを考え，これらの行ベクタを意味素性の数 N_s だけ縦方向に連接した次のような N_s 行 N_v 列の行列 SM を考える．これを意味素性行列と呼ぶ．

$$SM = \underbrace{\begin{bmatrix} s_1 & s_1 & \cdots & s_1 \\ s_2 & s_2 & \cdots & s_2 \\ \vdots & \vdots & \ddots & \vdots \\ s_{N_s} & s_{N_s} & \cdots & s_{N_s} \end{bmatrix}}_{N_v}$$

意味素性行列を結合価行列に左からかけて得られる N_s 行 N_p 列の行列を優先度行列と呼ぶ．ただし，意味素性同士の積は以下のように定義する．

$$s_i \times s_j = \begin{cases} 1 & \text{if } s_i = s_j \\ 0 & \text{otherwise} \end{cases}$$

すなわち、優先度行列 PM は、

$$SM \times VM = \begin{array}{c} \\ s_1 \\ s_2 \\ \vdots \\ s_{N_s} \end{array} \begin{array}{c} \begin{array}{cccc} p_1 & p_2 & \cdots & p_{N_p} \end{array} \\ \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1N_p} \\ n_{21} & n_{22} & \cdots & n_{2N_p} \\ \vdots & \vdots & \ddots & \vdots \\ n_{N_s1} & n_{N_s2} & \cdots & n_{N_sN_p} \end{bmatrix} \end{array}$$

優先度行列の各要素 n_{ij} は N_v 個の動詞の中で意味素性 s_i と格 p_j の組合せをとることのできる動詞の数を示している。したがって、優先度行列の要素の値が小さい場合には、その要素に対応する意味素性と格の組を結合価ベクタに含む動詞が少ないということなので、それだけ動詞を限定する力が強いことになる。逆に優先度行列の要素の値が大きい場合には、その意味素性と格の組は動詞を限定しにくいということになる。優先度行列が佐伯のいうかかりの広さに関する定量的な尺度を与えるというのが我々の主張である。したがって、後置詞句の語順を考えると、優先度行列の要素の値が大きい格と意味素性を持つ後置詞句ほど先行しやすいことが予想できる。

2.3 結合価行列の縮退

優先度行列は、辞書中の、あるいは、聞き手の意識にのぼっているすべての動詞に関して、どのような後置詞句が入力されやすいかという優先度を与える。これに対して、ある後置詞句が入力された後にどのような後置詞句が入力されやすいかを考えよう。たとえば、名詞句の意味素性が s_i、格助詞が p_j である後置詞句が入力された時点での優先度行列はどのようになるかを調べよう。この場合、格 p_j に意味素性 s_i をもたない動詞が文末に現れることはないので、結合価行列において p_j 列の要素に値 s_i を持たない行は削除することができる。次に p_j の意味素性が決定されたので、結合価行列の p_j に相当する列も削除することができる。この操作を結合価行列の縮退と呼ぶ。

縮退した結合価行列に意味素性行列を左からかけると、意味素性 s_i を持つ名詞句と格助詞 p_j からなる後置詞句が入力された時点での優先度行列が得られる。この優先度行列は後置詞句 (s_i, p_j) が入力されたという条件つきのかかりの広さを表している。以下、同様に新しい後置詞句が入力されるたびに、結合価行列の縮退と優先度行列の計算をおこない、次に入力される後置詞句の優先度を計算することができる。

3 実験

本章では、IPAL 動詞辞書の結合価パタンの情報を用いて優先度行列を計算し、佐伯の提案した語順傾向と比較することによって我々の推定モデルの有効性を検討する。

3.1 IPAL 動詞辞書

IPAL 動詞辞書 [3] は計算機による日本語処理のために作成された動詞辞書で、基本的な和語動詞 861 語に関する統語的、意味的情報を含んでいる。IPAL 動詞辞書の見出し語は次のような階層を持ち、見出し語数は、見出しレベルで 861、サブエントリレベルで 3370 である。

見出し > 同音意義語番号 > サブエントリ番号

IPAL 動詞辞書で扱う格は、ガ、ヲ、ニ、カラ、ヘ、ト、ヨリ、デ、Φ である。ここで、Φ は格助詞を必要としないはだか格を表す。本実験では、はだか格以外の 8 つの格助詞について考慮する。それぞれの格には、その格がとりうる名詞句の意味素性が割り当てられている。IPAL では 20 種の意味素性を設定している。

3.2 結合価行列と優先度行列

まず、IPAL 動詞辞書から見出し語、同音意義語番号、サブエントリ番号、結合価パタン情報を抽出する。結合価パタンの交替要素について結合価パタンを展開し、展開した結合価パタンに 0 から順に結合価パタン番号をふる。したがって、各動詞は統語的、意味的情報の違いも含めて、見出し語、同音意義語番号、サブエントリ番号、結合価パタン番号の 4 つ組で表現できる。ただし、ここでいう「動詞」は 2 章で述べたように統語的、意味的な違いも考慮したものである。こうして 8829 個の動詞が得られるが、ここで、ガ－ガ、ニ－ニのように同じ格を複数持つ結合価パタンを持つ動詞については結合価ベクトルが構成できないので削除する。これを除くと 8062 個の動詞が残る。

以上のようにして得られた 8062 の動詞から 2 章で述べたように、結合価ベクタ、結合価行列を構成する。IPAL 動詞辞書の意味素性の数は 20 なので 20 行

8062 列の意味素性行列を作り、それを結合価行列に左からかけると 20 行 8 列の優先度行列を得る.

3.3 格の優先順序

かかりの広さが広い、すなわち優先度行列の要素の値が大きいかかり要素ほど先行するという佐伯の仮説にしたがえば、かかりの広さという観点から各動詞について最適な後置詞句配置が決定できる. 各動詞について以下の手続きにより、最適な後置詞句の配列を計算する.

(1) 動詞を 1 つ取り出し、以下の処理を結合価パタンのすべての非 NON 要素についておこなう.

(2) その動詞の結合価パタンに含まれる後置詞句 (意味素性と格の組) の中から現在の優先度行列で最大の値を持つものを選択する.

(3) 選択した後置詞句に基づいて結合価行列を縮退し、新しい優先度行列を計算する.

(4) 新しい優先度行列と結合価パタン中の残りの後置詞句について (2) から繰り返す.

このようにして得られた各動詞の最適な後置詞句配列から 2 つの格の優先関係を計算する. たとえば、結合価パタンの最適配列は次のような形式で得られる.

443 28 HUM:ガ HUM:ニ ACT:ヲ

ここで、最初の数字は結合価パタンの識別番号、次の数字はその結合価パタンを持つ動詞の数、そして残りが後置詞句の配列である. この項目からは、ガ > ニ、ガ > ヲ、ニ > ヲという優先関係にそれぞれ 28 点ずつを与える. このようにして得られた結果からおおよそ次のような格の優先関係が得られた.

ガ > ヲ > ニ > カラ > デ

3.4 検討

IPAL の動詞辞書について我々のモデルが推定する語順傾向を佐伯の観察と比較した結果、(傾向 1) と (傾向 3) については強い相関が、(傾向 4) については弱い相関が観察できる. 特に (傾向 1) に関して、我々のモデルではニ格にどのような意味素性を持つ名詞句をとるかによって細かい分析が可能である点に注目して欲しい. 宮島は川端康成の新文章読本から抽出した文について語順傾向を調査している [1]. その結果、(傾向 4) と (傾向 5) を除き、佐伯とほぼ同様な傾向を観察している. 宮島の結果と我々の実験結果から (傾向 4) は必ずしも強い傾向ではないことが予想できる. より詳細な検討については、文献 [5] を参照して欲しい.

4 おわりに

本論文では佐伯の提案したかかりの広さという概念を用い、後置詞句の並びと動詞という構造を持つ日本語の単文の語順を推定するモデルを提案した. モデルの妥当性を検証するために、IPAL 基本動詞辞書から結合価パタンを抽出し、実験をおこなった. その結果モデルの予測する語順傾向が佐伯の分析に一致することを示した. このモデルの自然言語処理への応用としては、音声認識における認識結果の候補の絞り込み、名詞句の意味のあいまい性の解消、文生成における基本語順の決定など考えられる [4]. 本論文では、文脈の影響、省略語、係り助詞などについては考慮していないが、これらの現象を我々のモデルでどのように扱うかについては、文献 [4] を参照して欲しい.

参考文献

[1] 宮島達夫. カカリの位置. 計量国語学. Vol. 23. pp. 3-11. 1962.

[2] 佐伯哲夫. 現代日本語の語順. 笠間書院. 1975.

[3] 情報処理振興事業協会. 計算機用日本語基本動詞辞書. 1986.

[4] 徳永健伸、岩山真、乾健太郎、田中穂積. 日本語語順の推定モデルとその応用. 情報処理学会 自然言語処理研究会. Vol. NL81-2. . 1991.

[5] 徳永健伸、田中穂積. 結合価に基づく日本語語順の推定モデル. 計量国語学. Vol. 18. No. 2. pp. 53-65. 1991.

[6] 北原保雄. 日本語動詞の研究. 大修館書店. 1981.