

共起パターンと名詞意味素性のテキストからの抽出

7G-6

木子健一郎

納富一宏

内山明彦

早稲田大学理工学部

1. はじめに

現在、自然言語による文書の電子化が、盛んに行なわれているが、一方では自然言語の解析・生成に必要な解析用辞書の構築を、効率よく行なう方法が問題となっている。

特に自然言語の解析には、動詞の共起パターンと、格を決定するための名詞の意味素性に関する知識の存在が有用であることが知られている^{[1][2]}。しかしながら、人間の手でそれらを構築しようとする試みは、その量が膨大であるために、いまだ実用に至っていない。

そこで今回、共起パターン、および未知名詞の意味素性をテキストから抽出する方式を提案する。本方式ではテキストを解析し、未知名詞が無い場合には格助詞と名詞の意味素性を用いて共起パターンを抽出し、未知名詞が入っている場合にはその格を検出し、共起パターンを用いてその名詞に対して意味素性の付加を行なう。このような方式をとることににより、意味素性と共起パターンの両方を相補的に獲得することができ、知識の不備による解析の失敗を減少させることができると考えられる。

2. 共起パターンと意味素性

着目する動詞がどのような名詞とともに使われるかを表すのが共起パターンである。ここで共起パターンの表現方法には、表層的に格助詞のみを用いる結合価のレベルと、名詞が用言に対してどのような格として働くかを問題にする格支配のレベルとがある。ここでは、動詞と共起する名詞にかかる格助詞とその意味素性、および名詞が取る深層格の情報を共起パターンとして用いる。深層格には全ての動詞が取る、場所や時間などがはいる任意格を設ける。

解析に用いる深層格の例を表1. に、共起パターンの例を図1. に示す^[1]。

表1. 共起パターンに用いる深層格

動作主格 : AGENT	対象格 : OBJECT
道具手段格 : MEANS	終点格 : GOAL
起点格 : SOURCE	構成格 : MATERIAL
原因理由格 : REASON	任意格 : ARBITRARY

選ぶ: A [hum]が + O [div]を + G [con]に + V
 信じる: A [hum]が + O [hum]を + V
 : A [hum]が + O [con]を + V

図1. 共起パターンの例

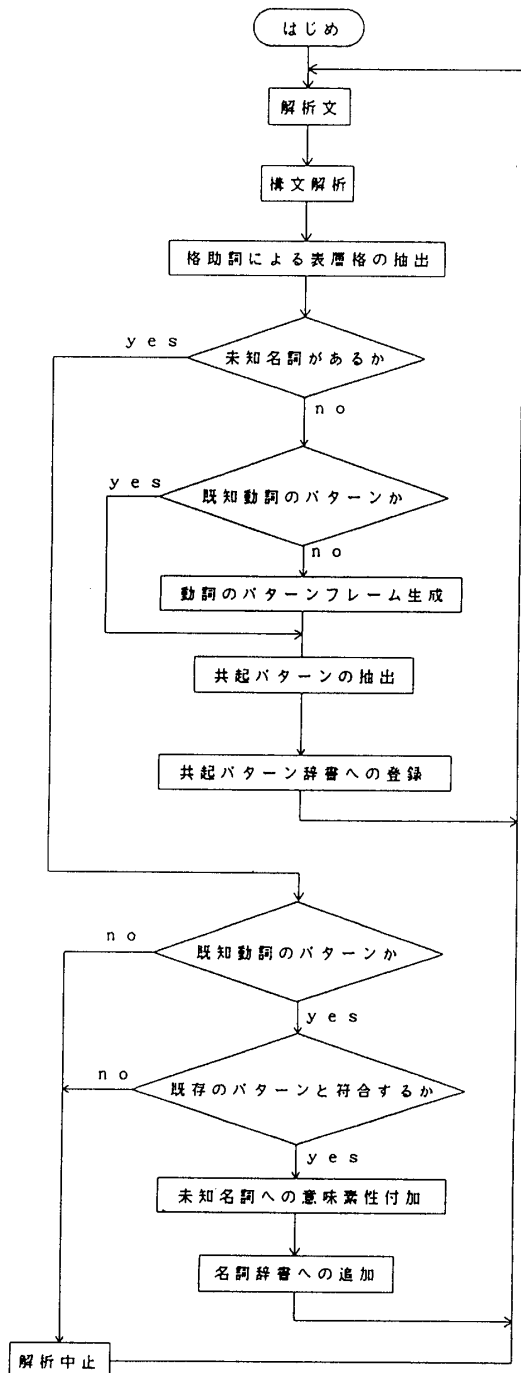


図2. 全体の処理手順

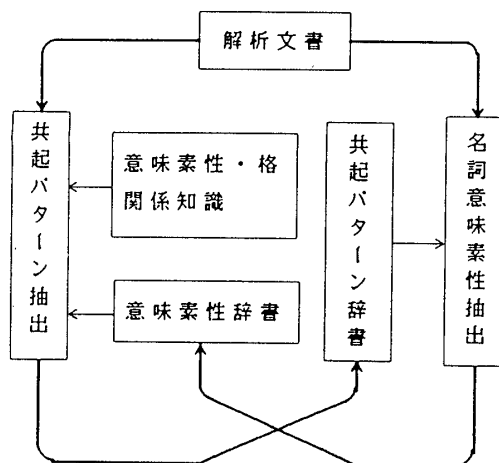


図3. システムのブロック図

3. 処理方法

全体の処理手順を、図2. に、システムのブロック図を、図3. に示す。

(I) 共起パターンの抽出

本方式では、テキストを構文解析し、格助詞と名詞の意味素性に関する制限知識をもとにして共起パターンの抽出を行なう。抽出は以下の手順で行なわれる。

- ① 格助詞を用いて獲得する共起パターンの候補を提出する。
- ② それぞれの格が取る意味素性の制限情報を用いて候補をしぼる。
- ③ 次に、既存の共起パターンとの照合を行ない、一致するか包含関係があればそれを採用する。
- ④ 一致しない場合、候補が一つであればそれを採用し記録する。候補をしぼりきれない場合には、人間側に提示して判断を仰ぐ。
- ⑤ 解析文の動詞が未知である場合には、その動詞のパターンフレームを生成し、パターン情報を付加する。

ここで、各名詞はその意味素性がすべて既知であるものとし、係受け等に曖昧性を残すものについては、誤った共起関係の抽出を防ぐため、処理の対象から除外する。図4. に例を示す。

この例文からは、表層的に格助詞を用いて a)、b) のような候補が抽出される。さらに、名詞の意味素性に関する制限情報を用いると、b) が選択できるが、この方法でも候補を特定できない場合には、人間が判断することになる。

例文：彼は屋台で昼食を食べた。

食べる：

- a) A [hum] が I [con] で O [con] を + V
b) A [hum] が A r [loc] で O [con] を + V

屋台： [l o c] → b) を抽出

図4. 共起パターンの抽出

(II) 共起パターンを用いた名詞意味素性の抽出

(I) で示した共起パターンの抽出法ではすべての格の意味素性が判っていることが前提である。しかしながら実際問題として、テキストから解析を行なう場合に未知の名詞がまったく無いということは考えられない。

そこで逆に、既知の共起パターンを利用して、未知名詞の意味素性の抽出を行なうことを考える。

処理手順を以下に示す。

- ① 既知の共起パターンと一致し、未知名詞を含む格が(ただ一つ)存在するかどうか判定する。
- ② 未知名詞に対して、概念フレームを生成する。
- ③ 共起パターン中の、該当する格要素の意味素性をパターン辞書より抽出する。
- ④ これを現在の未知名詞の意味素性として概念フレームに付加するが、複数存在する場合には、これらを提示して、判断させる。
- ⑤ この単語を名詞辞書に登録し、名詞辞書を逐次拡大する。

この手順によって未知名詞とその意味素性が獲得でき、以降の解析に役立てることができると考えられる。

この場合、未知の名詞が解析文中の格に、単独で存在することが条件であり、(I) と同様に係受けなどの曖昧な文については、解析を行なわない。

解析の様子を図5. に示す。

例文：議長が演説を試みた。

未知名詞：演説

共起パターン(既知)

試みる：A [hum] が + O [act] を + V

意味素性付加

演説：[a c t]

図5. 未知名詞に対する意味素性の付加

4. おわりに

以上、テキストから格情報と意味素性を持った共起パターンと、未知名詞に対する意味素性の抽出を行なう方式について、提案した。現在この方式に基づいた辞書獲得システムの製作を行なっている。今後は本システムを完成させ、実際に辞書の獲得を試みる予定である。

参考文献

- [1] 水谷静夫、他：朝倉日本語新講座3、文法と意味、朝倉書店
[2] 野村浩郷：自然言語処理の基礎技術、電子情報通信学会
[3] 長尾真：言語工学、昭晃堂