自然言語処理 141－3
(2001. 1. 25)

係り受け解析実験による動詞と格標識との多項関係共起知識の評価

嘉寿 毅　永井 秀利　中村 貞吾　野村 浩郷

九州工業大学 情報工学部 知能情報工学科
E-mail: {kanaga,nagai,teigo,nomura}@dumbo.ai.kyutech.ac.jp

語彙に関する最も基礎的な知識の一つに，動詞と格標識との共起知識がある．従来の多くの研究ではある動詞と個々の格標識との共起知識を二項関係と見なし獲得されていたが，格標識間の相関性を考えた場合，これでは十分な知識とは言えない．そこで，本研究では，動詞の用法のモデルを提案し，それに基づいてある動詞と格標識集合との共起頻度を推定することにより，多項関係としての共起知識を獲得することを目指す．本稿では，京都大学テキストコーパスを用いて共起知識の獲得を行った．さらに，その知識に基づく係り先推定実験により，本手法で獲得した知識の有効性を評価した．

Evaluation of Cooccurrence Knowledge as N-ary Relation between a Verb and Case Markers

Takeshi Kanaga, Hidetoshi Nagai,
Teigo Nakamura and Hirosato Nomura

Department of Artificial Intelligence, Kyushu Institute of Technology
E-mail: {kanaga,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp

One of the most basic information about vocabularies is the knowledge of cooccurrence between a verb and case markers. On many past works, the coocurrence knowledge has been acquired as unary relations between a verb and each case marker, but correlation of case markers has not been considered. Therefore, we propose a model of verb usage as n-ary relation, and based on it, we estimate strength of cooccurrence between a verb and each set of case markers. In this paper, we acquire the knowledge from Kyoto University text corpus, and evaluate it to disambiguate of case marker attachment.

1 はじめに

自然言語処理において，述語の意味を正しく理解することは重要なことである．述語の中でも動詞は意味，用法が広範であるため，動詞に関する語彙知識を正確に獲得することは言語処理において重要な課題の一つとなっている．

従来獲得が進められてきた多くの語彙知識は，動詞を中心とした格パターン知識であった．この格パターン知識は「ある動詞に対してどのような格要素がどの程度の頻度で共起しているか」といったことを捉えており，結果として，共起知識は動詞と格要素の二項関係に注目して獲得されていた．しかし，文を構成する格要素は集合としてその文の状況を示す構成要素群になると考えられる．なぜなら，述語と共起する格要素集合は述語を中心として相関関係をもち，文意を構成する要素集合となっていると考えられるからである．したがって，ある動詞と共起する格要素集合の関係，すなわち，格要素間の多項関係に注目し，共起知識の獲得を行う必要があると考える．多項関係としての共起知識を統計的に獲得する研究 ([2][3][4] など) では，格パターンの数が大きくなるため，膨大なサンプルを必要とする．しかし近い将来まで含め，そのような巨大なサンプルの存在を期待するのは非現実的である．

本研究では限られたサンプル集合を用いることを前提とし，格要素中の格標識に注目して，ある動詞と共起する格標識の出現の組み合せとその出現頻度を共起知識として獲得することを目的とする．動詞と共起する格標識集合は表層的な出現を捉えることとした．また，動詞を含む幾つかの助詞相当語句を格標識とし，助詞相当語句内の語が意味的・構文的に整合性のある係り受け関係を持つようにした．獲得した共起知識の妥当性を検証するため，係り受け解析実験を行った．

2 格標識

2.1 格標識の共起

格標識がある動詞に係る時，その動詞と格標識は共起していると言う．文がある状況を表す時，述語 (動詞) に係る格要素は一般的に複数現れると言える．図1に，動詞と格要素の係り受け関係の例を示す．

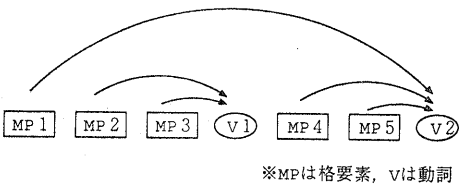


図 1: 動詞と格要素の共起

従来の多くの研究では，動詞と格標識の共起知識は1対1の二項関係として獲得されてきた．しかし，このような手法のもとでは，MP2とMP3のように係り先が同じ動詞である場合，MP2とMP3の間にどのような相関関係があるのか，また，その他の省略された可能性のある格標識に対してどのような頻度でどのような共起をするのかといった情報は得ることができない．格標識間の相関性に注目し，動詞と格標識の共起を多項関係と見なすことによって，ある動詞に対して共起する格標識の組み合せとその出現頻度を共起知識として獲得することができる．

例えば，同じように限定を意味する格標識「だけ・のみ」はある動詞に対して相反的に出現する傾向が強いと考えられる．一方で，出来事が起こる起点を表す格「から」と終点を表す格「まで」は同時に生起することも多いと考えられる．こういった動詞を中心とする格標識間の相関関係は，個々の格標識を単一のものとして捉えるのではなく，動詞を中心とした格標識の集合として捉えなければ，明らかにならない．

2.2 格標識と助詞相当語句

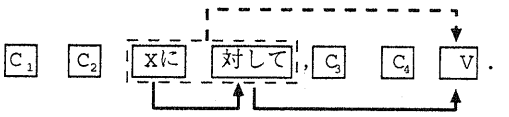
日本語において，表層的に見た格標識は助詞である．一般的には，次に示すような単体の助詞を格標識として認める．

- 単体の助詞
  に，が，は，で，と，も，を，まで，から，等

また，連続した助詞によって構成される助詞相当語句も格標識とすることもある．

- 連続した助詞による助詞相当語句
  には，では，からは，とも，とは，等

ここで助詞相当語句とは単体の助詞以外の語や語句が文節間の係り受けを決定する助詞の働きを持つものとする．上記の前提にそうと，「に対して」などの動詞を含む慣用表現も助詞相当語句とすることができる．このような動詞を含む格助詞相当語句は文節中の語の意味的・構文的な係り受けの強さが異なることが問題となる．慣用表現「X に対して」を例として，構文上の係り受け関係と意味上の係り受け関係の違いを図 2 に示す．



※ただし，cは文節，xは語，vは動詞

図 2: 慣用表現「に対して」の係り受け関係

図 2 では，構文上の X の係り先は太線の矢印で示すように動詞「対し」であるが，これは意味的に強い係り受け関係であるとはいえない．むしろ，語 X は文の主動詞である動詞 V と意味的に強いつながりを持っていると言える．しかし，点線の矢印で示すように語 X と動詞 V との関係は構文上，直接表現されていない．

ここで，「に対して」を格標識と見なすと，「X に対して」は一つの文節となり，語 X は点線の矢印で示すように動詞 V との間に係り受け関係を持つことが可能となる．その結果，意味的に強い関係を持つ語 X と動詞 V の間に構文的にも意味的にも整合性のある係り受けの関係が成り立つ．テキストコーパス上でこういった慣用表現 (動詞を含む助詞相当語句) は表層的に容易に取り出すことが可能であり，これらの助詞相当語句を格標識として認めることによって，意味的な係り受けの強さを係り受け関係に反映させることができる．また，動詞を含む助詞相当語句を格標識に含めることは，動詞の表現をより詳しく捉える基準を増やすことであると考える．例えば，助詞相当語句「に対して」を格標識に加える前では，慣用表現「に対して」も助詞「に」も格標識「に」という同じ表現基準によって動詞の用法を表現する．しかし，動詞を含む助詞相当語句を格標識に加えることによって，それらは独立に動詞の性質を表す基準となる．

3 動詞と格標識集合の共起

2.1 節で述べたように，従来の多くの研究では，動詞と格要素間の関係を単項関係であるとみなし，共起情報を獲得してきた．しかし，このような手法を用いて獲得された知識では，ある動詞に対してどのような格標識の組み合せがどの程度出現するのかは未知数である．そこで，動詞に対して格標識は集合単位で共起するという考えのもと，ある動詞と共起する格標識の組み合せとその出現頻度を共起知識として獲得する．

3.1 共起知識の獲得元

共起知識の獲得には，何らかのコーパスが知識元となる．現在，様々なコーパスが利用可能ではあるが，解析器等で自動獲得されたコーパスは解析の誤差を内包しているため，知識の獲得対象として充分に信頼性のおけるものとはいえない．一方で，自動獲得されたコーパスを人手で修正したコーパスも公開されているが，最終的に人間によるチェックが必要とされることから大規模なものはあまり存在しない．特に，統計的な手法を用いる本研究では格標識の出現パターンが膨大になるため，必要とされるサンプルの量は非常に大きなものになると考えられる．しかし，現時点でそれほど大きなコーパスを期待することは非現実的である．この様な前提のもとで，現在利用可能なコーパスを利用し，できるだけ動詞の性質をうまくとらえた知識を獲得することを目指す．本研究では，知識獲得元として京都大学テキストコーパス [5] を用いた．このコーパスには毎日新聞の記事，約 2 万文が含まれており，形態素・構文解析した結果に対して人手による修正がなされている．コーパスには文節の係り先に関する情報が記述されており，これらの情報を次章の共起知識の評価実験で用いる．

3.2 動詞の用法の捉え方

サンプルの類似性を推定する場合，多次元空間上の分布を捉える手法がよく用いられる．本研究における共起知識の獲得では，各格標識を軸とする多次元空間上でのサンプルの重心と分布傾向を共起知識とする．ここで，サンプルは一文中に格標識がそれぞれ幾つ現れるかというデータを持つこととする．このような多次元空間上でのある動詞に対するサンプルの分布は，

最もよく使われる格標識の組み合せが最も重心の近くに分布し，重心から離れるにしたがってサンプルの出現頻度が減少することが予想される．このような多次元空間において，サンプル間の類似性を取り出すためには"距離"という尺度をその基準とする．上記の通り，重心の最も近くに分布するサンプルは，その動詞の用法を最もよく表す格標識の組み合わせであると考えられる．また，サンプル間の距離が短いということは類似性が高いと言うことを意味する．

本研究における多次元空間上の分布の傾向をうまく捉えるには，分布の分散や方向性を考慮し，サンプルを楕円体上に捉える必要がある．このため，最も一般的な距離の概念であるEuclid距離や格子点上に存在するデータを対象としたManhattan距離などでは，本研究におけるサンプルの分布傾向をうまく取り出すことはできない．上記のような距離概念では格標識間の相関性を取り扱うことはできず，格標識間の関係は二項関係として獲得されると言える．よって，本研究では類似性基準として次章に示すMahalanobis距離を使用する．

3.3 Mahalanobis距離

分布の分散方向を考慮にいれた距離概念の一つにMahalanobis距離がある．以下にMahalanobis距離を求める式を示す．

$$D_M(x)^2 = (x-m)^t C^{-1}(x-m) \quad (1)$$

※x はサンプル，m はサンプルの重心
C は分散・共分散行列

本研究における共起知識とはサンプルの重心と分散・共分散行列を求めることと等しい．分布の重心に近いほどサンプルは動詞のよく使われる用法を表していると言える．また，分散・共分散行列はサンプルの分布傾向を表していると言える．

Mahalanobis距離を求める課程では分散・共分散行列の逆行列を求める必要がある．しかし，単純にサンプルから分散・共分散行列の逆行列を求めようとした場合，$|C| = 0$ となるため逆行列が求まらない．原因としては以下の2つがあげられる．

- ある格標識の分散値が0となる．これは，ある格標識が全てのサンプルにおいて必ず同じ数だけ出現することを意味する．(0個を含む)
- 相関係数が1，または-1となる．これは，ある格標識集合の出現が完全に同時であるか，または完全に排他であるかということを意味する．

今回の共起知識獲得材料である京都大学テキストコーパスには様々な各パターンを獲得するのに充分な量の文が含まれているとは言いえない．しかし，このような状況下でサンプルの分布傾向をうまく捉えるために以下のような対処法を用いる．

- 分散値0を避けるために，各座標値が0.5となるダミーデータを加える．
- 相関係数の絶対値を1より小さくするため共分散成分に(N - 1)/Nを掛ける．

ある動詞に対してある格標識が全く出現しないとその格標識を軸とするサンプルの分散値が0となる．本稿では動詞を含む助詞相当語句を格標識として追加しているが，特にこの様な格標識の出現頻度は単体の助詞や助詞の連続による助詞相当語句に比べきわめて低い．このような分散値0を回避するためにはあるサンプルが他のいかなるサンプルとも異なった値を持つ必要があり，各座標値が0.5となるダミーデータを加えた．相関係数に関しては，その絶対値が1となる場合が問題であると考え，その絶対値を1より小さくするために(N - 1)/Nをかけることとする．実際には，ある格標識の組合せが完全に排他的に，または，同時的に出現することはありうると考えられる．格標識の相関性を調査し，同一の性質を持つ格標識に対しては座標軸の縮退をすることが可能であるが，本稿では格標識の選定に関する議論はしないこととする．上記のような対処方法をとることによって，数学的には正しくないものの，より多くのサンプル数を含むほど獲得する知識に対するサンプル数の影響を小さくすることが可能である．

3.4 共起知識の獲得対象

コーパスより能動形での出現頻度の高い動詞を56個抜きだし，共起知識の獲得対象とした．受動形，使役形については文型が変化するため，今回は除外している．次にコーパス中で抽出し

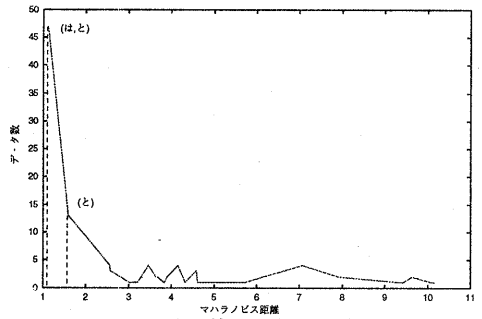
た動詞が含まれる文中に出現する助詞群を抜きだし格標識集合とした．また，連続して出現する助詞の中で個々の助詞としての働きと連続した助詞としての働きが異なると考えられるものは独立した格標識とした．助詞と同様に文節間の係り受けを決定する働きを持つ動詞を含む助詞相当語句についても独立した格標識とした．共起知識の対象とした格標識は図3の38個である．

単体の助詞	
に，が，は，で，と，も，を，まで，から	9個
連続した助詞	
には，では，にも，とも，とは，からは	
にまで，までは	8個
助詞と動詞による助詞相当語句	
に対して，に関して，に向けて，等	21個

図 3: 共起知識獲得の対象となる格標識

3.5 共起知識の獲得

獲得した共起知識がどのようにサンプルの分布を捉えているかを動詞「話す」におけるサンプルデータの分布を用いて説明する．



※ (は，と) 等は，格標識集合

図 4: 動詞「話す」におけるサンプルの分布

図4では，重心 (Mahalanobis 距離が0) の近くに多くのデータが集まり，重心からの Mahalanobis 距離が大きくなるにつれて，サンプルの出現頻度が減少する分布となっている．これは，重心近くに動詞の最も使われやすい用法を表す格標識が集中し，重心からの距離が離れるにつれて出現するサンプル数が少なくなるという傾向が現れており，動詞の用法的な特徴をうまく捉えていると言えよう．しかし，次の図5

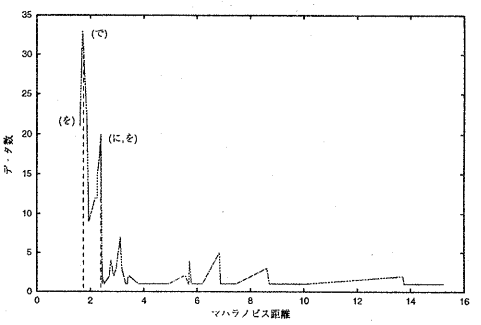


図 5: 動詞「開く」におけるサンプルの分布

のように重心から最近傍の格子点がサンプル数最大となっていない動詞の分布も幾つか見られた．これは表層的な格標識集合に対して，動詞の用法が複数現れているからであると考えられる．動詞の用法が一つしか現れていないサンプルの分布では，重心から最近傍の格子点上のサンプル数が最大になり，それ以降は重心から離れるにつれて出現頻度が減少していくと予想されるが，図5の分布はそれとは異なり，ピークとなる点が複数現れている．このような動詞の用法を分類 [1] することによって，同じ動詞でも異なった用法に対するサンプルの共起知識を獲得することができる．

4 共起知識の評価実験

第3章では，テキストコーパスより共起知識を獲得する手法を示した．獲得された共起知識の有用性を検証するために，格要素の係り先の推定実験を行う．格要素の係り先に複数の候補が現れる場合を係り受け推定実験の対象とするため，文中に出現する動詞が1個の場合や格要素集合の係り受け解候補数が1個の場合は実験対象に含めないこととする．

本研究では動詞を含む助詞相当語句を格標識に含めたため，その前後の解析精度を比較した．次に，文中の係り受け精度をより詳しく調べるために，格要素ごとの係り受け推定精度を求めた．第3章では，獲得した共起知識に対する動詞の用法分類を紹介したが，このような知識を利用すると解候補の優先順位付けに対して

重み付けを行うことができると考えられる．しかし，サンプルデータが少ない動詞が幾つか存在したため，全ての動詞に対する統一的な分類が行なえず，本稿では動詞の用法を用いた係り受けの推定は行っていない．したがって，共起知識としてのサンプルの分布の重心や分布傾向を捉えている分散・共分散行列などが，ある動詞に対して純粋に用法ごとに獲得されていない状況で実験を行った．

4.1 実験の手順

係り受けの推定を行う実験手法について以下に説明する．実験材料は京都大学テキストコーパス [5] である．格要素の係り先となる動詞は共起知識獲得実験と同様の 56 個とした．これは，獲得された共起知識を用いて解候補に優先順位付けを行うためである．コーパスより正しい係り受けの解を抽出し，Mahalanobis 距離を用いて優先順位付けされた解候補と比較することよって係り受け推定実験を行う．次に実験の手順を示す．

※ただし，以下の説明では v:動詞，f:最初の動詞に係る格要素，s:2 番目の動詞に係る格要素，p:格要素，-:それ以外の要素（文節）を示す．

＜係り受け推定実験手順＞

1. まず，実験対象として一文中に 2〜4 個の動詞を含む文を抽出する．実際には一文中に最高で 6 個まで動詞が含まれる場合があるが，サンプル集合が小さいために信頼性のおける解析結果が得られないと考え，本稿では実験の対象外とした．動詞を含む助詞相当語句の構文的・意味的な係り受けの整合性をとるため，本稿では 21 個の動詞を含む助詞相当語句を格標識としてのその働きを見る．動詞を含む助詞相当語句を格標識に加える場合，格標識中の動詞は動詞とみなさない．

2. 次にテキストコーパスより，手順 1 で抽出したデータ中の格標識の係り先を求め，係り受け関係の正解データとする．文中には，述語として動詞以外に形容詞，形容動詞，サ変動詞の名詞型などが存在するが，動詞以外の述語に係る格要素は実験対象とせず，削除した．なぜなら，本実験では動詞と格標識集合の共起知識に注目しており，形容詞や形容動詞等，動詞以外の述語に対する係り受けは本実験の趣旨に反すると考えたからである．サ変動詞の名詞型は動詞とみなすことによって係り先を推定することができると考えられるが，本実験では対象外とした．動詞が 2 個の時の，正解データの例を次に示す．

   - s - f v1 - - s v2

3. 手順 2 で作成した正解データを用いて，係り受けの解候補を作成する．ここで，「文節はそれ以前の文節には係らない」という規則と，係り受け非交差のヒューリスティックを用いて，現実には存在し得ない解候補を削除する．係り受けが交差する場合は実際には存在するが，その数が少ないと考え，解候補には含めなかった．次に上記の正解データから作成された解候補を示す．

   1) - f - f v1 - - s v2
   2) - s - f v1 - - s v2
   3) - s - s v1 - - s v2

4. 手順 3 で求めたそれぞれの解候補に対して Mahalanobis 距離を用いて優先順位付けを行う．ここで，第 3 章で獲得した共起知識を利用する．まず，ある動詞に注目して解候補ごとに格要素の集合に対する重心からの Mahalanobis 距離を求める．

   動詞 $v1$ − − $d_{v1}(1), d_{v1}(2), d_{v1}(3)$
   動詞 $v2$ − − $d_{v2}(1), d_{v2}(2), d_{v2}(3)$

   次に，各解候補ごとに Mahalanobis 距離の 2 乗和をとる．2 乗和をとる理由は，大きく分布の重心から離れた格標識集合の共起に対し，重み付けを行うためである．この操作によって，ある動詞に対して極端に重心から離れた格標識集合が存在すれば，他の格標識集合がどのような共起をしても，解候補の優先順位を低くすることができる．

   $1) d_{v1}(1)^2 + d_{v2}(1)^2$
   $2) d_{v1}(2)^2 + d_{v2}(2)^2$
   $3) d_{v1}(3)^2 + d_{v2}(3)^2$

   最後に，Mahalanobis 距離の小さいものから解候補に優先順位を付けを行う．

4.2 異なる格標識集合に対する解析結果

格標識間を多項関係とみなすことよって共起知識の獲得材料となるコーパスの規模は非常に大きなものを必要とする．特に，本研究で格標識とみなす動詞を含む助詞相当語句は著者の言い回しなどに影響を受けやすいため，文中に出現する頻度が低い．このような条件下において，動詞を含む助詞相当語句を格標識として加える前後の係り受け解析精度がどのように変化するかを調べた．なお，格標識は第3.4章の図3に示すものである．

4.2.1 格標識17個の時

以下に示す表は正解データが何番目に優先順位付けされた解候補と一致するかを示している．また，比較は一文単位で行っているため，係り受け推定の対象となる格要素すべての係り先が正しく推定されていなければ正解数にカウントされない．

優先順位	動詞数		
	2個	3個	4個
1位	70.4%	52.7%	41.3%
2位	22.7%	18.3%	12.9%
3位	4.9%	9.1%	6.5%
4位	1.0%	5.2%	6.5%
5位	0.2%	4.6%	2.6%
それ以上	0.2%	8.7%	27.6%
係り受け交差	0.6%	1.4%	2.6%
対象データ数	1213	562	155

表1: 格標識数17個の時の一文単位の係り受け解析精度

表1より，1位に優先順位付けされた解候補が最も正解データと一致する割合が高く，優先順位が下がるにつれて正解率が下がっていることが分かる．本実験は表層的に格標識を捉えており，荒い解析ながらも上位に正解データが集中していることが分かる．次に文中に出現する動詞ごとの解候補数を示す．

動詞数	2個	3個	4個
解候補数の平均	2.79	9.24	31.0

表2: 格標識数17個の時の係り受け解候補数

表2では，一文中の動詞数に対し，係り受けの解候補数がどのように増加するかを示している．一般的に，文中に現れる動詞数が増えると，係り受け非交差のヒューリスティック等を適応しても，動詞数に比例して解候補は指数倍に増加するといえる．その結果，動詞数の増加にともなって解候補数が極端に増えていることが図2より分かる．最後に，格要素単位の解析精度を示す．

動詞数	2個	3個	4個
解析精度	81.8%	80.9%	80.6%
対象格要素数	2108	1692	609

表3: 格標識数17個の時の係り受け精度

4.2.2 格標識38個の時

格標識に動詞を含む助詞相当語句を認めた場合の結果を次に示す．

優先順位	動詞数		
	2個	3個	4個
1位	60.6%	42.0%	26.7%
2位	29.3%	21.9%	12.3%
3位	6.2%	10.8%	10.3%
4位	0.7%	6.5%	8.9%
5位	0.2%	4.2%	5.5%
それ以上	0.3%	9.7%	28.1%
係り受け交差	2.7%	4.9%	8.2%
対象データ数	1014	526	146

表4: 格標識数38個の時の係り受け解析精度

表1と表4より，1位に優先順位付けされた解候補の精度を比較すると，格標識が38個の時の方は一文中のそれぞれの動詞数における推定実験において10%ほど精度が低下していることが分かる．次に，文中に出現する動詞ごとの解候補数の平均と文節単位の解析精度を示す．

動詞数	2個	3個	4個
解候補数の平均	2.69	7.28	20.73

表5: 格標識数38個の時の係り受け解候補数

動詞数	2個	3個	4個
解析精度	74.9%	71.9%	66.8%
対象格要素数	1612	1256	422

表6: 格標識数38個の時の文節単位の係り受け精度

表2と比較すると，動詞を含む助詞相当語句を格標識に加えた後の解候補数の平均は加える前に比べ，減少していることが分かる．

4.3 評価と考察

表層的に格標識の出現をとらえ，意味解析を行わない中で，全体として良好な結果が得られ

たと考え，本研究における共起知識獲得手法の有効性が確認した．また，実験材料である京都大学テキストコーパスは人手による修正が行われているため，量的に見て大きなものとは言えないが，この様な比較的規模の小さなコーパスからもうまく共起知識を獲得できていると言える．次に本実験における共起知識獲得手法についての考察を示す．

まず，動詞を含む助詞相当語句を格標識に加えた場合，加える前と比較すると全体的に係り受けの推定精度の低下が見られる．その理由として，まず，組合せパターンの増加によるサンプル数の不足が原因として挙げられる．対処方法としては機械的に形態素・構文解析された結果を用いて，大量のサンプルの中から動詞を含む格標識に対する共起知識を獲得し，その結果を係り受け解析に応用する事が挙げられる．解析器の誤差による誤った知識も本稿で紹介した動詞の用法のモデル化によってある程度吸収できると考えられる．次に，単純に格標識数を増加させたことによって同等の使われ方をされるべき格標識が別のものとして扱われるようになったことも原因として挙げられる．その結果，分布が空間上で薄く広がってしまい，分布の重心にサンプル数が集中すると言う特徴が強く現れなくなっていると考えられる．この問題に対しては，格標識の働きを整理し，同等の働きをする格標識を同一化することによって解決できると考える．

本実験では共起知識獲得段階において複数の用法があると考えられるサンプルの分布に対し，用法ごとの分類を行っていない．このような動詞の用法を完全には反映していない結果を係り受け推定に利用していることも精度を低下させていると原因と言える．例えば，動詞の用法が2個出現していると考えられる場合，分布全体の重心とそれぞれの用法の重心は一般的に異なっており，また，全体の分布傾向とそれぞれの用法に対する分布傾向も一致していない．ある動詞におけるそれぞれの用法をクラスタリングし，その結果を用いることによって，動詞のそれぞれの用法に対する重心と分布傾向を利用した解析が行えると考える．ただし，各用法の出現頻度は一般的には異なると考えられるため，各用法でのサンプルの出現頻度に応じた重み付けを行う必要があろう．

5 まとめ

本稿では，ある動詞に対して共起する格標識を集合として捉え，動詞と格標識集合間の共起知識を獲得した．共起知識の対象としては，単体の助詞や，連続した助詞によって構成される助詞相当語句に，動詞を含む助詞相当語句を加えた．このことによって，格要素内の語の意味的・構文的係り受け関係に整合性を持たせ，動詞の表現する内容をより多角的にとらえた．共起知識獲得手法の評価実験として，構文解析の代表的な研究課題である係り受け解析を表層的に行った．

実験では，一文中に動詞が2〜4個現れる文を対象として動詞を含む助詞相当語句を格標識に含める前後の推定精度を一文単位，格要素単位でそれぞれ求めた．その結果，動詞を含む助詞相当語句を格標識に加えた後で係り受け推定精度の低下が見られたものの，全体的には良好な結果を得ることができ，本研究における共起知識の獲得手法が有用であることを確認した．今後の課題としては動詞の用法をクラスタリングすることによって，動詞のそれぞれの用法に対する知識を獲得することが挙げられる．

参考文献

[1] 永井 秀利，中村 貞吾，野村 浩郷：多項関係としての格標識共起知識の獲得，情報処理学会研究報告 2000-NL-136, pp. 63-70 2000

[2] 宇津呂 武仁，宮田 高志，松本 裕治：最大エントロピー法による下位範疇化の確率モデル学習および統語的曖昧性解消，情報処理学会研究報告 97-NL-119-11, pp. 69-76 1997

[3] 宮田 高志，宇津呂 武仁，松本 裕治：Bayesian Network による下位範疇化の確率モデルおよびその学習，情報処理学会研究報告 97-NL-119-12, pp. 77-84 1997

[4] 福本 文代，佐野 洋，斎藤 葉子，福本 淳一：係り受けの強度に基づく依存文法，情報処理学会論文誌 Vol.33, No.10, pp. 1211-1223(1992)

[5] 足立 顕，牧野 武則：表層格と動詞の関係に基づく動詞の自動分類，情報処理学会研究報告 2000-139-13 pp. 93-100 2000

[6] 黒橋禎夫 他：京都大学テキストコーパス 作業マニュアル，京都大学テキストコーパス Version2.0 付属ドキュメント 1998

[7] まつもとゆきひろ/石塚圭樹 共著，オブジェクト指向スクリプト言語 Ruby :アスキー出版者

[8] 奥津 敬一郎 他：いわゆる日本語助詞の研究，凡人社