4H−1

動詞と格標識の多項関係知識に基づく動詞の用法のクラスタリングとその評価

嘉壽 毅† 永井 秀利† 中村 貞吾† 野村 浩郷†

九州工業大学 情報工学部†

1 研究の背景と目的

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．

我々は「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，格要素間の多項関係に注目し，共起知識の獲得を行ってきた．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．多項関係としての共起知識を統計的に獲得する研究では，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り，限られたサンプル集合より有用な共起知識を獲得する必要がある．獲得した共起知識の中には動詞の用法が複数現れていると考えられるため，動詞の用法ごとにサンプルのクラスタリングを行う必要がある．

本研究では，限られたサンプル集合から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．また，獲得した共起知識中に現れる動詞の用法をクラスタリングし，動詞の用法に基づいた知識を獲得することを目的とする．最後に，獲得した共起知識の妥当性を検証するために係り受け推定実検を行った．

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．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，係り先が同じ格標識があった場合，その格標識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．

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，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も格標識とする．本研究では次に示す単体の助詞と連続した助詞を格標識として用いる．

- 単体の助詞 (計10個)
  に，が，は，で，と，も，を，まで，から，では
- 連続した助詞による助詞相当語句 (計7個)
  には，にも，とも，とは，からは，にまで，までは

格標識として扱うべき表現は他にも存在するが，サンプル中の出現頻度が極めて低いものは対象外とした．

3 動詞と格標識集合の共起

3.1 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において，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．空間上において，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サンプル間の類似性が高いと言うことを意味する．一般的によく利用される距離としてはEuclid距離やManhattan距離があるが，このような距離の概念では本研究で獲得を目指す格標識間の相関関係を考慮することができない．よって本研究では，サンプルの分布の軸と分散を考慮した距離の1つであるMahalanobis距離を採用する．

3.2 共起知識の獲得対象

共起知識の獲得元は京都大学テキストコーパスver.2.0を用いた．共起知識獲得対象となる動詞はコーパスより能動形での出現頻度の高い動詞を56個抜きだした．ただし，受動形，使役形については文型が変化するため，今回は除外した．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．ここで，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．

4 動詞の用法のクラスタリング

動詞は一般的に多義性を持つと考えられる．前章で獲得した共起知識中には，ある動詞に対して複数の用法を持つと考えられるサンプルの分布が幾つか見られた．このように複数出現した動詞の用法に対してはクラスタリングを行うことでそれぞれの用法の重心を中心としたサンプルの分布を獲得することができる．

本研究におけるクラスタリングは参考文献 [1] の手法を基本とした．クラスタ解析手法は様々なものがあるが，本研究ではサンプルが格子点上に存在するといった特殊な分布であるために，一般的な手法を用いてクラスタリングを行うことは難しい．クラスタリングの基本的な方針は次に示すようなものとなる．

- 構成するクラスタは，3.1 章で述べたように，重心付近にサンプルが集中し，重心からはなれるにつれてサンプルの出現が疎らになる傾向を持つ．
- 生成されるクラスタ数は限定しない．最初からクラスタ数を限定すると分離されるべき用法が他の用法に吸収されたり，逆に，不必要な分離が発生する可能性を持つからである．(それぞれの格子点が 1 つのクラスタに成るような極度な細分化は避ける．)
- クラスタは隣接する (Manhattan 距離が 1) 格子点を吸収し，距離が 2 以上の格子点は他のクラスタを構成することとする．

上記の方針に従って動詞の用法のクラスタリングを行った結果，動詞'いる'，'見える'，'聞く'，'なる'，'残る'で，大きなクラスタを 2 つ生成した．その他の動詞は 1 つのクラスタのみか，または，大きなクラスタと格子点ごとに構成されるような小さなクラスタに分かれた．

5 係り受け推定実検

獲得した共起知識の有効性を検証するため，係り受け推定実検を行った．実験の手順として，まず，一文中に複数の動詞 (共起知識獲得の対象となる 56 個の動詞) を含む文を抽出し，係り受け可能な解候補を作成した．ここで，一文中の動詞が 5 個以上出現するものはサンプル数が少ないため実験の対象外とした．

推定実験はクラスタリングを行う前の共起知識を用いた場合と，クラスタリングによって動詞の用法ごとに獲得された共起知識を用いた場合に対して行った．以下に文節単位での推定精度を示す．

動詞数	対象文節数	分類前	分類後
2 個	2108	81.8%	86.6%
3 個	1692	80.9%	85.0%
4 個	609	80.6%	82.6%

表 1: 分節単位の係り受け推定精度

表 1 は，優先順位付けで 1 位となった解候補中の格標識の係り先が係り受け正解データと一致する数をカウントした．表 1 より，一文中の動詞数にかかわらず，動詞の用法のクラスタリング結果を用いることによって推定精度が向上していることが分かる．次に，一文単位での推定精度を示す．

優先順位	分類前			分類後		
	2 個	3 個	4 個	2 個	3 個	4 個
1 位	70.4	52.7	41.3	77.7	61.3	49.1
2 位	22.7	18.3	12.9	19.0	17.2	20.0
3 位	4.9	9.1	6.5	3.2	10.0	6.3
4 位	1.0	5.2	6.5	0.1	4.4	8.9
5 位	0.2	4.6	2.6	0.0	2.0	4.5
6 以上	0.2	8.7	27.6	0.0	4.9	11.2
データ数	1213	562	155	1213	562	155

表 2: 文単位の係り受け推定精度 (単位は%)

表 2 では，優先順位づけされた解候補が一文単位でどの程度正解データと一致しているかを示している．文節単位の推定精度向上を反映して，動詞数に限らず一文単位の係り受け精度も向上が見られる．しかしながら，一文中の動詞数が増加すると推定精度が減少している．

6 まとめと考察

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を用いた．その後，獲得した共起知識を動詞の用法に注目してクラスタリングを行った．獲得した共起知識と動詞の用法のクラスタリング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い，動詞の用法に注目してクラスタリングされた共起知識の有用性を検証した．

実験では，一文中に動詞が 2〜4 個現れる文を対象とし，動詞の用法をクラスタリングした前後の共起知識を利用して係り受けの推定精度を一文単位，各要素単位でそれぞれ求めた．その結果，クラスタリング後の共起知識を用いることによって推定精度が一割程度向上し，全体的に良好な結果を得た．精度向上の理由としては，動詞の用法を反映した共起知識を用いることによって，用法毎のサンプルの重心と分散傾向を利用した優先順位付けが行われたことが上げられる．今後の課題としては動詞の用法のクラスタリング時に 1 つの格子点毎にクラスタを構成するような極度のクラスタの細分化解消が挙げられる．また，解候補の優先順位付けを行うときに，クラスタ毎にどのような重みづけを行うかも課題となる．

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] 嘉寿 毅，永井 秀利，中村 貞吾，野村 浩郷：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 2001-NL-141, pp. 13-20 2001