語順を考慮した格フレームの提案と獲得手法

大竹 清敬[†]　根津 雅彦[†*]
増山 繁[†]（正員）　山本 和英[††]

Automated Acquisition of Case Frames with Case Order

Kiyonori OHTAKE[†], Masahiko NEZU[†*], *Nonmembers*,
Shigeru MASUYAMA[†], *Member*, and
Kazuhide YAMAMOTO[††], *Nonmember*

[†] 豊橋技術科学大学知識情報工学系，豊橋市
Dept. of Knowledge-based Information Eng., Toyohashi University of Technology, Toyohashi-shi, 441-8580 Japan
[††] ATR 音声翻訳通信研究所，京都府
ATR Interpreting Telecommunications Research Laboratories 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto-fu, 619-0288 Japan
[*] 現在，(株) 名鉄コンピュータサービス

あらまし　語順を考慮した格フレーム獲得のための格遷移ネットワークモデルを提案する．このモデルに対し，予備的検討のための実験を行った．その結果，語順を考慮した格フレーム獲得に有効であり，実用に際してはより高精度な意味素辞書が必要であることもわかった．

キーワード　格フレーム，語順，自動獲得，コーパス

1. まえがき

日本語における格フレーム獲得について述べる．本研究の特徴は語順(注1)を考慮して格フレームを獲得するところにある．本論文では，格フレーム獲得のために，語順を考慮したモデルを提案し，その予備的検討を行ったので報告する．

機械翻訳や文章理解などの自然言語処理システムにおいて，辞書が重要な役割を果たしている．辞書にどれだけの記述が含まれているかが，システムの到達可能な性能を規定する．そのため，自然言語処理システムの高度化は必然的に豊富な辞書記述を要求する．

そのような辞書の一つとして格フレームがある．自然言語処理において，格フレームは文の意味を表現するために必要不可欠なものとして取り扱われてきた．また，その必要性から格フレームの自動獲得に関する研究が数多く行われてきた [1], [3], [4], [6], [9]．

従来の日本語を対象とした格フレームの自動獲得に関する研究は，格要素の出現順序（以下語順とする）について全く考慮していない．本研究では格要素の出現順序に着目して，単一言語コーパスから統計的に格フレームを獲得するためのモデルを提案する．

(注1)：本研究では，格要素の出現順序を指す．

語順の変化は格フレームが表現する意味に大きな変化を与えない．しかしながら，語順情報は語用論的文脈解析において，強調などの情報を抽出する上で役立つことが期待できる．更に，大量のコーパスを用いて統計的に学習を行うため，統計的に自然な語順を示すことが可能となる．日本語は語順が自由な言語だといわれることがあるが，最も自然な順序は次のようになることが多いようである：「時の成分–所の成分–主格–与格–対格–動詞」[8]．そして，これは統計的にも確かめられている [2]．しかし，文献 [2] にて調査された語順は動詞を無視した一般的なものであり，語順の変化のパターンが動詞によって異なるかどうかについて，調査及び言及されていない．

動詞によって語順変化のパターンが異なるのであれば，それは動詞の意味分類を詳細化する上で役立つ情報となるはずである．したがって，語順を考慮した格フレームを獲得することによって，語順情報が動詞の意味分類を詳細化する能力を有するかどうかを検討できる．また，語順情報を保持した格フレームによって，一般的な語順の言語生成を行うことが可能となる．更に，強調が必要な場合には，強調として自然な語順を動詞ごとに提供できる．

本研究では，1 文中に一つの動詞をもち，動詞の後に名詞が存在しない文を単文とする．また，名詞とそれに伴う格助詞をまとめて格要素と呼ぶが，本研究では名詞に意味素性を割り当てるため，意味素性と格助詞の組を格要素とする．

2. 格遷移ネットワーク

この章では，本研究で提案する格遷移ネットワークモデルについて説明する．格フレーム自動獲得の手法として様々な方法が考えられるが，本研究ではコーパス上の事例から統計的に獲得することを考える．そのような言語モデルとしては，従来 n-gram が頻繁に用いられてきた．しかしながら，n-gram モデルを使用する際，n が大きくなるに従って，必要なコーパス量が大きくなる．また，十分なコーパスを用意できたとしても，モデルの内部状態数が爆発する可能性がある．そこで，本研究では，格要素列を格要素の bi-gram で近似し，格遷移ネットワーク上に配置することにより，状態数を抑えて格要素列を保持し，出現位置情報も保持できるモデルを提案する．

格遷移ネットワークモデルは直観的には図 1によって示される．この図において，$s_{i,m}$ は格要素に対応し，弧は出現順序を示している．また，i は単文内におけ

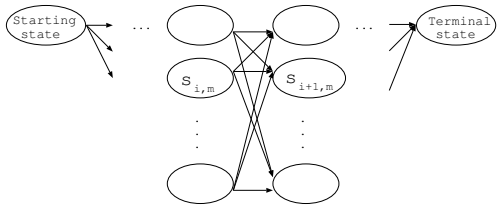

図 1　格遷移ネットワークモデル
Fig. 1　A case transition network model.

る格要素の位置を示しており，m は格要素を示している．このモデルを用いてある単文について弧の重み付け（以下，学習と呼ぶ）を行うときの概要は以下のとおりである．

（1） 文の先頭から順に格要素を調べる．

（2） 次の格要素が出現すると，状態遷移し，その弧の重みを計算する．

（3） 動詞の出現によって，最終状態へ遷移したならば，その単文についての学習は終了する．

学習の結果，初期状態から最終状態までの値をもつ弧による路が，格フレームを表す．

格遷移ネットワークの定義を以下に示す．格遷移ネットワーク $N(v,n) = (S, A, w)$ は，ある動詞 v とその格要素の数 n が与えられたときに，状態の集合 S，弧の集合 A，並びに，弧の重み w から構成される．

[状態] 初期状態，中間状態，最終状態の 3 種類がある．初期状態が単文の文頭，中間状態が格要素，最終状態が動詞にそれぞれ対応する．初期状態，最終状態は格遷移ネットワーク上にそれぞれ一つずつ存在する．中間状態はその出現位置によって区別され，一つの格要素が一つの中間状態に対応するので，中間状態は出現位置ごとの層をなす．中間状態を $s_{i,m}(1 \leq i \leq n)$ と表記する．ここで，i は単文における格要素の出現位置を示し，m は格要素を示す．$i = 0$ となる状態は初期状態のみであり，$i = n + 1$ となる状態は最終状態のみである．

[弧] 状態 $s_{i,k}$ から状態 $s_{i+1,l}$ への遷移を弧 $(s_{i,k}, s_{i+1,l})$ によって表現する．また，弧 $(s_{i,k}, s_{i+1,l})$ は重み $w(s_{i,k}, s_{i+1,l})$ をもつ．弧の向きによって語順を表現し，弧の重みによってその語順の起こりやすさを表現する．

[弧の重み] 弧 $(s_{i,k}, s_{i+1,l})$ の重み $w(s_{i,k}, s_{i+1,l})$ を，条件付き確率 $p(s_{i+1,l}|s_{i,k})$ で与える．$p(s_{i,k}, s_{i+1,l})$ を状態 $s_{i,k}$ から状態 $s_{i+1,l}$ への遷移確率とし，$p(s_{i,k})$ を状態 $s_{i,k}$ の出現確率とするとき，$p(s_{i+1,l}|s_{i,k})$ は (1) 式によって求められる．

$$p(s_{i+1,l}|s_{i,k}) = \frac{p(s_{i,k}, s_{i+1,l})}{p(s_{i,k})} \quad (1)$$

格要素の並びである格要素列は，格遷移ネットワークにおける初期状態から最終状態へ向けて弧をたどったときの路で表現される．格遷移ネットワークにおいては，格要素列の出現確率をその路の上の弧の重みの積によって推定する．

3. 実　　験

格遷移ネットワークモデルの妥当性を検討する実験を二つ行った．これらの実験によって格遷移ネットワークモデルがどのような特性を有するのかを明らかにする．格遷移ネットワークモデルにおいては格要素列を格要素の bi-gram で近似しているため，格要素列間において，その出現確率の大小関係が崩れる可能性がある．これを調査するために，一つ目の実験として学習コーパスにおける格要素列の頻度とその格要素列の格遷移ネットワークによる推定出現確率の値を比較する実験を行った．また，本来の目的である語順情報の保持に関しても，既に述べた理由により，出現頻度と矛盾する語順を推定する可能性がある．そのため，二つ目の実験として，これがどの程度あるのか調査を行った．

実験の対象としたコーパスは日本経済新聞の 1990 年から 1992 年の CD-ROM 版である．コーパスから単文を抽出した結果，557048 文が得られた．

本手法では，名詞に意味素性を割り当てなければならない．意味素性は [7] に記載されている，18 種類の素性を用いた．また，名詞に素性を割り当てるために角川類語新辞典 [5] の各カテゴリに素性を割り当てて，人手によって一部を修正，追加したものを意味素性辞書とした．本研究で対象とした格助詞は「は，が，を，に，から，へ，と，より，で」の 9 種類である．本来，「は」は格助詞ではないが，ガ格または，ヲ格に伴って助詞「は」が出現する場合は，助詞「が」または「を」は削除されるため，「は」から，その格を推定できる可能性がある．しかしながら，実際にその推定を行うには多くの問題を伴うため，本研究では「は」を特別な助詞として格助詞と同等に扱うこととした．

次に，ある動詞に対する格要素列の抽出方法を以下に示す．

（1） あらかじめコーパスを JUMAN (注2) によって形態素解析しておき，指定された動詞と格要素数をもつ単文を得ておく．

（2） KNP(注3) を用いて単文を構文解析する．

（3） 構文解析結果から，動詞にかかる文節のうち，名詞と格助詞によって構成されているものを格要素とする．名詞は接尾辞も含めて名詞とする．

（4） 格要素内の名詞に意味素性を割り当てる．このとき，完全一致によって割り当てることができなかった場合，後方からの最長一致で割り当て可能な素性を探す．もしそのような素性があれば，それをその名詞の意味素性とする．最終的に割り当てが不可能であった場合，割り当て不可能の素性とする．また，1 つの名詞に複数の意味素性が割り当てられる場合もある．

（5） 格要素列を出力する．

抽出された格要素列に対して，解析失敗などの理由により極端に頻度が小さい格要素列がある．これを次の手法によって削除する．

（1） 文内の語順を無視した格助詞の列（以下 C とする）を考え，その頻度を計算する．

（2） C の頻度が $\frac{\text{文の総数}}{\text{語順を無視した格助詞列の数}} \times \frac{\text{しきい値 (\%)}}{100}$ よりも小さければ C は無効であるとして削除する．

（3） すべての語順を無視した格助詞列に対して上の 2 つを実行する．

今回の実験ではしきい値を 15%として削除した．

更に，一つの名詞に対して複数の意味素性が割りあてられている場合は，一つの格要素列から意味素性の数に応じた格要素列を作成する．こうして得た格要素列を学習用格要素列とする．

実験の対象とした動詞は，コーパス中で出現頻度が大きい上位 5 個「開く，始める，まとめる，出る，入る」を選択した．

3.1 推定出現確率と格要素列の出現頻度との関係

動詞「開く」の格要素数 3 について実験した推定出現確率と格要素列の出現頻度との関係を図 2に示す．また，他の動詞についても図 2と同様に右上りの結果が得られた．

調査した格要素数を 3 としたのは，次の二つの理由による．一つは，格遷移ネットワークの性質から，2 以下では，学習格要素列と推定格要素列との間に差がないためである．もう一つは，IPAL の基本動詞辞書 [7]

(注2)：http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html
(注3)：http://pine.kuee.kyoto-u.ac.jp/nl-resource/knp.html

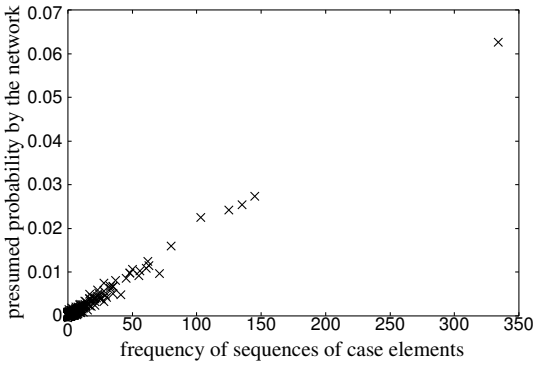

図 2 動詞「開く」の推定出現確率と格要素列の出現頻度との関係
Fig. 2 Relation between presumed probabilities by the network and sequences of case elements in the corpus for the verb, '*hiraku*(to open)'

表 1 単文における格要素数の構成
Table 1 Component ratio for the number of case elements in a simple sentence

格要素数	1	2	3	4	5
必須格のみの場合（%）	15.0	64.4	20.5	0.2	0
任意格を含む場合（%）	6.8	50.5	36.7	5.4	0.5

中において動詞がとり得る格要素数を調査したところ，2 の次は 3 が最も大きいためである．なお，表 1にこの調査結果を示す．また，格要素数は最大 5 であった．

3.2 語順情報に関する調査

学習後の格遷移ネットワークが語順情報をどの程度保持しているかを評価するために，格要素数を 3 に限定して調査を行った．

まず，任意の格要素列 α_i に対して，α_i の語順を入れ替えてできる格要素列は，格要素数が 3 の場合，合計六つできる．そして，α_i を除く残りの列による集合を $\{\alpha_{i1}, ..., \alpha_{i5}\}$ とする．ここで，更に，

$$g(\alpha_i) \equiv \begin{cases} 1, & \forall j,\ p(\alpha_i) > p(\alpha_{ij}), \\ & s.t.\ f(\alpha_i) > f(\alpha_{ij}), \\ 0, & \text{otherwise}, \end{cases}$$

となる関数 g を定義する．ただし，$f(\alpha_i)$ は α_i の出現頻度を表し，$p(\alpha_i)$ は α_i の推定出現確率を表すものとする．ここで異なり学習格要素列集合を L とすると，語順保持率は次式で定義される．

$$\text{語順保持率} \equiv \frac{\sum_{\forall \alpha_i \in L} g(\alpha_i)}{|L|} \times 100(\%) \qquad (2)$$

動詞「開く，始める，まとめる，出る，入る」について語順保持率を調べた結果を表2に示す．

表2　語順保持率（格要素数3）
Table 2　Preservation ratio of case orders.(3 case elements)

動詞	語順保持率（%）	対象となった単文の数
開く	95.0	2143
始める	98.7	1304
まとめる	97.2	205
出る	97.6	336
入る	95.7	419
平均	96.8	881.4

4. 考　　察

図2から，格遷移ネットワークが学習格要素列の出現に比例した推定出現確率を提示することがわかる．そのため，格遷移ネットワークは学習格要素列中に出現した格要素列について，その頻度に比例した推定出現確率を出力すること，また，格遷移ネットワークが期待したように働くことがわかる．次に，語順保持率は，表2に示した結果となっており，いずれの動詞に対しても高い結果を示した．このことから，格遷移ネットワークは語順情報を十分に保持したモデルであることが結論づけられる．

今回の実験では，格遷移ネットワークの有効性を示すために，基礎的な実験のみを行った．しかし，最終的に格フレームを獲得するためには，推定出現確率値に対してしきい値を設けるなどして，妥当な格フレームを出力しなければならないし，また，そのために必要なコーパス量などを調査しなければならないが，それらの検討は今後の課題である．

格遷移ネットワークは既に述べたように，ある意味でデータスパースネスに対応しているが，データスパースネスに対する有効性を今回の実験では検証することができなかった．その原因の一つに意味素性辞書の多義性が挙げられる．名詞に意味素性を割り当てる際に，複数の素性が割り当て可能な場合には，可能な素性をすべて割り当てる．その結果，素性の数だけ格要素列が作成される．そのため，学習用格要素列の数がコーパス中の単文数に比べてかなり大きくなってしまう（例えば，動詞「開く」の格要素数3の場合，単文の数が2143に対して学習用格要素列の数は5256となる）．この事実が，格遷移ネットワークの学習に大きな影響を与えていることは容易に想像できる．これらのことから，格遷移ネットワークによる格フレーム推定の精度を向上させるためには，更に質のよい意味素性辞書を用いること，並びに，多義性の解消が必須である．これらの課題に対してどう取り組んでいくかは今後の課題である．

5. む す び

本論文では，語順を考慮した格フレーム獲得のために格遷移ネットワークモデルを提案した．格遷移ネットワークは語順情報を保持して格フレームを獲得できるのが特徴である．今回の実験では，提案した格遷移ネットワークが学習に用いた格要素列の出現頻度を反映した推定出現確率を出力すること，及び，語順情報を保持する能力が十分にあることを確認した．今後いくつかの自然言語処理システムを用いて，語順情報の必要性及び有効性について検討していく予定である．

謝辞　本研究で，シソーラスに使用した「角川類語新辞典」[5] を機械可読辞書の形で提供頂き，その使用許可を頂いた（株）角川書店に深謝する．また，本研究の一部は，文部省科学研究費特定領域研究B（2）の援助を受けて行われた．

文　　献

[1] M.R. Brent. “Automatic acquisition of subcategorization frames from untagged text,” Proc. of the 29th Annual Meeting of the ACL, pp. 209–214, 1991.
[2] 国立国語研究所, 現代雑誌九十種の用語用字 (3)–分析–, pp. 171–239, 秀英出版, 1964.
[3] T. Mine, Masaru Higashi, and M. Amamiya. “Case frame acquisition and verb sense disambiguation on a large scale electronic dictionary,” Proc. of NLPRS ’97, pp. 221–226, 1997.
[4] 大石　亨, 松本裕治. “格パターン分析に基づく動詞の語彙知識獲得,” 情処学論, vol. 36, no. 11, pp. 2597–2610, 1995.
[5] 大野　晋, 浜西正人. 角川類語新辞典, 角川書店, 1981.
[6] 田中英輝, “動詞訳語選択のための「格フレーム木」の統計的な学習,” 自然言語処理, vol. 2, no. 3, pp. 49–72, 1995.
[7] 情報処理振興事業協会技術センター (IPA), 計算機用日本語基本動詞辞書 IPAL（Basic Verbs） –解説編–, 1987.
[8] 寺村秀夫, 鈴木　泰, 野田尚史, 矢澤真人 (編), ケーススタディ 日本文法. おうふう, 1987.
[9] T. Utsuro, Y. Matsumoto, and M. Nagao, “Verbal case frame acquisition from bilingual corpora,” Proc. of the IJCAI-93, Vol. 2, pp. 1150–1156, 1993.

（平成11年9月27日受付）