コーパスからの格要素列獲得における多義性への対応

大竹 清敬*　　増山 繁　　山本和英
豊橋技術科学大学 知識情報工学系　　ATR 音声言語通信研究所
kohtake@slt.atr.co.jp　masuyama@tutkie.tut.ac.jp　yamamoto@slt.atr.co.jp

1　はじめに

単一言語コーパスから格フレームとして用いることが可能な格要素列を獲得する手法について述べる．自然言語処理において，格フレームの重要性は認識されているものの，量，質ともに十分な格フレーム辞書が整備されているとはいい難い．その理由のひとつは，これまでに整備されてきた格フレーム辞書 [1, 2] が主に人手によるためである．人手による辞書整備では，大規模な辞書作成には膨大なコストがかかり，さらに，一貫性および客観性を保つことが困難である．

計算機およびネットワークの発展により，大量のコーパスを計算機上で容易に入手，処理できるようになった．そのため，大量のコーパスから格フレーム辞書を自動獲得する試みがいくつか行われてきた [3, 4]．格フレームの自動獲得における問題は主に次の 2 点である．(1) 統語的曖昧性．(2) 意味的曖昧性（特に名詞の多義性）．

宇津呂ら [3] は，二言語対訳コーパスを用いることで，これらの問題に対処している．また，Mine ら [4] は EDR コーパスとその解析結果である辞書を用いている．しかしながら，大量の二言語対訳コーパスを用意すること，ならびに，最新の解析済み辞書を整備することは高価である．そこで，本研究では，容易にしかも大量に入手可能な単言語のテキストコーパスから格フレームを獲得することを試みる．

我々が最終的に獲得しようとしている格フレームは，ある動詞がどのような格要素を，どのような語順でとりうるかを記述したものである [5]．本研究における格要素は名詞と格助詞により構成される．格要素列とはある動詞について，その動詞に係る格要素を出現順に並べたものである．

テキストコーパスから格フレームを獲得する試みとして既に，河原らによる研究 [6] がある．しかし，河原らは動詞とその直前の格要素の組による格フレームを考えており，格要素列を考えている本研究とは異なる．

動詞と名詞の関係を記述した格フレームを獲得するためには，構文解析（係り受け解析）が必要となる．しかし，一般に構文解析器によって出力された解析結果には誤りが含まれるため，これが格フレーム獲得において障害となる．また，名詞に意味素性を割当てて抽象化した意味素格フレームを得る場合は，意味素性を割当てる際の多義性が問題となる．

本研究では，対象とする文を単文に限定することにより，統語的曖昧性を抑える．さらに，意味素格フレームを作成するために，統計的に多義性を解消して名詞に意味素性を割当てる手法を提案する．

2　格要素列の抽出

この節では，コーパスからの格要素列抽出について述べる．コーパスから抽出された格要素列を，意味素性を割当てたものと区別するために，用例格要素列と呼ぶ．用例格要素列抽出の概要は次のようになる：(1) コーパスから単文を抽出．(2) それぞれの単文を構文解析．(3) 構文解析結果から文末にある動詞と，それに係る格助詞を持つ文節を格要素として格要素列を抽出．以下でそれぞれの処理について説明する．

単文の抽出 統語的曖昧性を避けるために，単文のみを対象とする．本研究における単文の定義は「動詞を 1 つだけ有し，動詞より後に名詞が存在しない文」とする．ただし，ここで言う動詞は形態素解析結果におけるもので，文末に存在することを想定している．この動詞の他に用言相当句 (サ変名詞+読点，形容詞+接尾辞など) が，動詞より前にあった場合は，単文と認定しない．また，単文の抽出には，形態素解析器のみを用いる．

構文解析 得られた単文をそれぞれ構文解析する．

格要素列の抽出 各文の構文解析結果から，(動詞，格要素数，格要素列) の 3 つ組を抽出する．

3　意味素性付与と多義性への対応

前節で得られた用例格要素列を，そのまま自然言語処理で用いることは，データスパースネスの問題により非

*現在，ATR 音声言語通信研究所

現実的である．そのため，一般に意味素性を割当てる．現在利用可能な意味素性辞書がいくつか存在するが，語彙の豊富さから，日本語語彙大系 [1] の意味コードを意味素性の近似として用いる．日本語語彙体系からの意味コードの取得にあたっては，NTT 日本語形態素解析ライブラリ ALTJAWS Ver.2.0 を使用する．

3.1 意味素性付与

日本語語彙大系の単語体系（以下では，意味素性辞書と呼ぶ）を用いることにより，名詞へ意味素性を割当てる事が可能となる．しかし，対象とする名詞が複合語などである場合は，その名詞がそのまま意味素性辞書へ記載されているとは限らない．そこで，本研究では，対象とする名詞に対して後方からの最長一致を行う．たとえば，意味素性を割当てる名詞として「ダボス会議」を考える.「ダボス会議」が意味素性辞書に存在しない場合は「ボス会議」→「ス会議」→「会議」の順に意味素性辞書を調べる．そして「会議」が意味素性辞書に存在した場合,「ダボス会議」にその意味素性を付与する．

ある文字列が意味素性辞書に存在する場合，その文字列が複数の品詞を持つ場合がある．意味素性辞書には次の 8 つの品詞がある.「名詞，サ変名詞，数詞，時詞，代名詞，固有名詞，接尾辞，接頭辞」そして，本研究における，複数品詞をもつ文字列に対する優先順位は，ここに挙げた順に低くする．これは，経験的にこのように定めた．たとえば，ある文字列が意味素性辞書に存在し，サ変名詞と接頭辞の項目を持つ場合は，サ変名詞を採用する．ただし，対象とする文字列が短かくなった場合は，接尾辞がふさわしい可能性が大きくなるので，接尾辞を優先する．具体的には，元の名詞の文字列長が 3 以上で，部分文字列を調査している時に部分文字列の長さが 2 以下になった場合は，接尾辞を最優先する．元の文字列長が 2 の場合は部分文字列長が 1 になると，接尾辞を最優先する．

用例格要素列内の各名詞に対して意味素性を割当てたものを素性付き用例格要素列と呼ぶ.

3.2 多義性への対応

ひとつの名詞に対して意味素性辞書に記述されている意味素性（意味コード）は多くの場合複数存在する．たとえば，名詞「単語」の意味素性は「1084 語」のみであるが，名詞「タンゴ」の意味素性は「1060 舞踊，1675 舞踊・演劇・諸芸，1055 楽曲」である．コーパスからの格要素列獲得にあたっては，この多義性の問題に対処しなければならない．

意味素性辞書のある項目に複数の意味素性が記述されている場合，その並びは，基本的なものからより派生的なものへの順になっている．そこで，多義性解消の単純なアプローチとして，複数の意味素性が記述されている場合は，最も基本的 (意味素性辞書の記述では最も左) なものを選択する方法がある．そして，我々がこのアプローチの予備的検討を行った結果，この方法による多義性解消の正解率は 7 割程度であった．

しかしながら，大量に用意した素性付き用例格要素列において，動詞と格を限定した場合，意味素性の出現に関して統計的差異が生ずると考える．そこで，次の 2 つの方針に基づき手法を検討した．(1) 意味素性辞書における素性の並びを多義性解消に利用する．(2) 多義性が生じていない事例を利用する．これらの方針に基づき，容易に得られる統計的情報とヒューリスティックスを用いた手法を考案し，多義性の解消を試みた．

まず，多義性解消のために必要な 3 種類の情報を素性付き用例格要素列から求める．

1. 多義性が生じなかった意味素性とその頻度：素性付き用例格要素列から，動詞 (v) と格助詞 (m) を限定した上で，単一の意味素性が割当てられている名詞を調べ，その素性 (f_i) の頻度 $f(v, m, f_i)$ を求める．

2. 多義性が生じた場合の共起頻度：素性付き用例格要素列において複数の意味素性 ($f_i,\ i = 1...$) が割当てられている名詞について，動詞 (v) と格助詞 (m) を限定した上で，それらの素性 ($f_i,\ f_j\ (i \neq j)$) の共起頻度を求める．たとえば，ある名詞に「f_1, f_2, f_3」の 3 つの意味素性が割当てられているとき，f_1-f_2, f_1-f_3, f_2-f_1, f_2-f_3, f_3-f_1, f_3-f_2 が 1 度ずつ共起している．これを $f_{co}(v, m, f_i, f_j)$ とする．さらに，動詞と格助詞が決められている場合に，ある意味素性 f_i と共起する他の意味素性の種類数を $N_{co}(v, m, f_i)$ で表わす．

3. 意味素性の統計的な重要度：動詞と格を限定した場合における意味素性の統計的な重要度を決定する．任意の動詞を v，格助詞を m，意味素性を f_i とするとき，f_i の重要度 $W(v, m, f_i)$ を次の式により求める．

$$\sum_{ce \in S(v,m,f_i)} \sum_{j=1}^{N(ce)} \frac{1}{p(ce_j, f_i)^3 \cdot N(ce)^{1.5}} \times \frac{1}{N_{co}(v, m, f_i)} \quad (1)$$

ここで，$S(v, m, f_i)$ は，動詞が v である素性付き用例格要素列において，格助詞 m を持つ格要素のうち

意味素性 f_i を含む全ての格要素に対応する．つまり，$S(v, m, f_i)$ は多重集合となる．

ce は任意の格要素を表している．ce によって示される格要素のうち，対応する名詞の文字列および品詞が同一でも複数の項目を持つものがある[1]．たとえば「なまず」は名詞であるが,「鯰」に対応する項目 (543 魚，842 魚介類) と「癜」に対応する項目 (2419 病気類) の 2 つがあるので，名詞「なまず」は 2 つの項目を持つ．また，ce が持つ項目の数を $N(ce)$ で表わす．ce に含まれるそれぞれの項目を $ce_j(1 \leq j \leq N(ce))$ に対応させる．

$p(ce_j, f_i)$ は項目 ce_j における意味素性 f_i の位置に対応する．たとえば，上述の例を用いると「鯰」の項目が「543 842」であるとき $p(「543\ 842」, 842) = 2$ となる．

直観的には，式 (1) によって示される重要度 W は，項目内での位置が後にいくほど (より派生的になるほど) そして，さまざまな他の意味素性と共起しやすければしやすいほど小さくなる．$N(ce)^{1.5}$ で割っているのは，ce が複数項目を持つ場合に，一つの名詞に関して意味素性を複数回割当ててしまうことを補正するためである．

そして，動詞 v の素性付き用例格要素列において，格助詞が m である格要素が複数の意味素性を持つ場合は次の処理を順次適用し，唯一の意味素性を決定する．

1. 格要素が複数の項目を持つ場合は，個々の項目に分割する $(ce_j(1 \leq j \leq N(ce)))$.
2. 個々の項目内の複数の意味素性 $(f_i(i = 1...) \in ce_j)$ それぞれについて，$f(v, m, f_i)/p(ce_j, f_i)$ を求める．
3. 上で求めた $f(v, m, f_i)/p(ce_j, f_i)$ の最大値が 1 以上ならば，その f_i を対象としている格要素の意味素性とする．
4. 個々の項目内の複数の意味素性 $(f_i(i = 1...) \in ce_j)$ のうち $W(v, m, f_i)$ が最大の f_i をその格要素の意味素性とする．
5. これまでの処理によって意味素性が一意に決定されなかった場合は最初の項目の最も基本的な意味素性 (辞書において最も左に記述されているもの) を対象としている格要素の意味素性とする．

以上の処理によって多義性が解消された素性付き用例格要素列を意味素格要素列という．

4 実験

前節で説明した多義性解消手法を計算機上へ実装し，実験を行った．使用したコーパスは日本経済新聞 CD-ROM 版の 90 年から 96 年である．使用した形態素解析器は JUMAN[2]，構文解析器は KNP[3]である．対象とした格助詞は,「が，を，に，から，へ，と，より，で」の 8 種類である．格助詞の選択にあたっては，KNP が IPAL の動詞辞書 [2] を格フレーム辞書として利用できることを考慮して，IPAL 動詞辞書に準拠した．ただし，格助詞が明示されないが，KNP の解析結果において，係り先が動詞で，かつ，＜時間＞＜数量＞＜係:無格＞の文節パターンを持つ文節は時を表す無格の格要素として扱う．なお，実験では，出現頻度の大きさから動詞「開く」を持つ格要素列を対象とした．

まず，90 年から 94 年のコーパスから単文を抽出し (924326 文)，KNP を用いて構文解析を行った．動詞「開く」を持つ用例格要素列 (9000 文) を抽出した．この格要素列に対して，意味素性の割当てを行ったところ，のべ 130 箇所の名詞について意味素性の割当てができなかった．この素性付き用例格要素列から $f(v, m, f_i)$, $f_{co}(v, m, f_i, f_j)$，$N_{co}(v, m, f_i)$，$W(v, m, f_i)$ を求めた．

次に，95 年のコーパスから動詞「開く」を持つ単文 (1705 文) を抽出し，同様に素性付き用例格要素列を求めた．このとき，のべ 30 箇所の名詞について意味素性割当てができなかった．割当て不可能となる原因の多くは，意味素性辞書に固有名詞の記述がないことである．そこで，90 年から 95 年に対する意味素性割当て結果を参考に可能なかぎり意味素性辞書の拡充を行った．同時に，90 年から 94 年のコーパスから得た統計的情報を基に，95 年のコーパスから得た素性付き用例格要素列の多義性解消を行い，提案手法のパラメータなどを調節し，最終的に 3 節で説明した値とした．

意味素性辞書拡充後，90 年から 94 年のコーパスに対する素性割当ができなかった箇所は，のべ 32 箇所，95 年のコーパスに対しては，のべ 16 箇所となった．

提案手法を評価するために，まず，96 年のコーパスからこれまでと同様に，動詞「開く」を持つ素性付き用例格要素列 (2058 文) を取得し，多義性解消手法を適用した．このとき，意味素性の割当てに失敗した箇所は，のべ 28 箇所であった．次に，多義性解消法のベースラインとして，多義性が生じた場合には，意味素性辞書において最も基本的な意味の意味素性を採用する手法を考える．このベースライン手法と提案手法を比較した．

多義性解消を行った格要素列から提案手法によって多義性が解消された格要素を持つ格要素列を任意に 90 個

[1] これは文字列照合に対応させるためにひらがなの項目を大量に用意したことと，名詞の品詞が本来は細分化されていたものを名詞にまとめあげた事に起因する．

[2] http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[3] http://pine.kuee.kyoto-u.ac.jp/nl-resource/knp.html

表 1: 多義性解消評価結果

手法	正解	不正解	正解率 (%)
提案手法	94	25	78.99
ベースライン手法	84	35	70.59

選び，多義性解消法の妥当性を評価した．この際，次の 4 つの条件のいずれかに該当する格要素は評価の対象から除いた．(1) 構文解析誤りにより，誤って格要素となっている．(2) 意味素性割当てに失敗しており，該当する意味素性がない．(3) 意味素性割当てに成功しているものの，意味素性辞書の項目に該当する意味素性が記述されていない．(4) 意味的に非常に近い意味素性が複数与えられており，いずれの意味素性を選んでも支障がない．その結果，90 個の格要素列における 119 個の格要素が多義性解消の評価対象となった．

提案手法によって多義性解消と，ベースライン手法による多義性解消の評価結果を表 1 に示す．また，提案手法が多義性を解消した際に用いた情報：(A) 単一意味素性の頻度 ($f(v, m, f_i)$) と，(B) 統計的重要度 ($W(v, m, f_i)$) の内訳を表 2 に示す．

表 2: 提案手法で用いた統計的情報の内訳

情報	正解	不正解	合計	正解率 (%)
情報 (A)	79	21	100	79.00
情報 (B)	15	4	19	78.95

5 考察

表 1 から提案手法は多義性の解消において効果があると言える．表 2 から多義性解消に用いた情報の多く (100/119) は単一意味素性の頻度である．このことから，単一意味素性の頻度が十分に得られるならば，この情報だけで 8 割程度 (79/100) の正解率を得られる．しかしながら，現実的にそのような条件を満たすことは困難である．実験では，単一意味素性の頻度だけでは判断できなかった 19 の格要素について，本研究で導入した統計的重要度によって約 8 割 (15/19) の正解率で多義性を解消することができた．

不正解事例のうち単一意味素性の頻度による場合の傾向として，判断に用いた値が十分な大きさを持っていないことが挙げられる．提案手法では，試行錯誤し，経験的にパラメータを決定した．しかしながら，提案手法における最適なパラメータの値は，コーパスの量，対象とする動詞などによって変動すると考えられる．これらの点に対する対処は今後の課題である．

6 むすび

テキストコーパスから格要素列を獲得する手法について述べた．格要素列を自動獲得する際に問題となる統語的曖昧性は，単文を対象とすることで抑えた．一般的な意味素性を持つ格フレーム獲得のために，格要素列に対して意味素性を割当てる手法を検討し，統計的情報に基づいて名詞の多義性を解消する手法を提案した．提案手法とベースライン手法とを比較した結果，多義性解消の正解率が 8%程度向上し，約 79%の正解率を達成した．今後，さまざまな動詞について実験と評価を繰り返す必要がある．

謝辞

本研究においてコーパスの研究目的での使用を許可して頂いた日本経済新聞社に感謝する．日本語語彙大系からの意味コード取得のための NTT 日本語形態素解析ライブラリ ALTJAWS Ver.2.0 の使用を許可して頂いた日本電信電話株式会社に深謝する．

参考文献

[1] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦（編）：日本語語彙大系 CD-ROM 版, 岩波書店 (1999).

[2] 情報処理振興事業協会技術センター：計算機用日本語基本動詞辞書 IPAL（Basic Verbs）–解説編– (1987).

[3] 宇津呂武仁, 松本裕治, 長尾眞：二言語対訳コーパスからの動詞の格フレーム獲得, 情報処理学会論文誌, Vol. 34, No. 5, pp. 913–924 (1993).

[4] Mine, T., Higashi, M. and Amamiya, M.: Case Frame Acquisition and Verb Sense Disambiguation on a Large Scale Electronic Dictionary, in *Proceedings of NLPRS '97*, pp. 221–226 (1997).

[5] 大竹清敬, 根津雅彦, 増山繁, 山本和英：語順を考慮した格フレームの提案と獲得手法, 電子情報通信学会論文誌, Vol. J-83-DII, No. 3, pp. 1060–1063 (2000).

[6] 河原大輔, 黒橋禎夫：用言と直前の格要素の組を単位とする格フレームの自動獲得, 情報処理学会研究報告 2000-NL-140, pp. 127–134 (2000).