モダリティおよび用言のガ格情報を付与したコーパスの作成

信本 浩二　木下 恭子　黒橋 禎夫
京都大学大学院 情報学研究科
{nobumoto,kinosita,kuro}@pine.kuee.kyoto-u.ac.jp

1 はじめに

これまでの言語処理の中心的課題は構文解析であったが，構文を明らかにするだけでは文の理解にはほど遠い．文の理解に近づくために最も必要なことの一つとして，日本語の場合には格解析がある．日本語では，格要素の語順の入れ替わりと省略，提題助詞や連体修飾節に関連する格助詞の欠落などの問題があり，構文解析を行っただけでは文中の「誰が何をどうした」という関係が明確にならない．この問題に対して，我々はタグ情報なしのコーパスから自動学習した格フレーム辞書によって格解析を行う方法を考案した [4]．これによって実用レベルのカバレッジの格解析が一応実現できたと考えている．

本稿では，その次の段階の問題として，文のモダリティの明確化と文中の用言のガ格の明確化の問題を扱う．日本語の文は，いわゆる内容表現の部分をそれ以外の部分 (非内容表現) が覆うような形で形成されている．この非内容表現の部分には，話者，文中の登場人物などの意図に関係する，情報伝達の観点から非常に重要な情報が含まれており，対話処理，要約処理などではこれらの正確な扱いが必要不可欠となる．

このような非内容表現について，これまでは，狭い意味でのモダリティ(真性モダリティ) について多くの研究がなされてきたが，実際の文中にはそれだけではなく様々な非内容表現が存在し，それらの広い意味でのモダリティ表現を網羅的に扱うことが本稿の目的である．

本稿で扱うもう一つの問題は，用言のガ格の明確化の問題である．機械翻訳システムは，すでに広く一般に利用されるソフトウェアとなってきたが，その翻訳の質はまだ十分なものとはいえない．翻訳がおかしい場合を調べてみると，その多くは格要素の省略，特にガ格の省略に関連する問題であることが多い．日本語の場合，ガ格の省略は非常に頻繁に起る現象であるが，機械翻訳に限らず，自然言語アプリケーションにおいて，その正確な扱いは最も重要な課題であるといってよい．

基本的にすべての用言がガ格を持つと考えられるから，格解析によって構文的に直接つながっているものの中にガ格の格要素がないことがわかれば，ガ格が省略されていると考えて，文脈からガ格を推定する必要がある．ここで重要なことは，用言に付随するモダリティ表現がガ格の推定と深く関連するということである．一つには，モダリティ表現からガ格が話者 (筆者) であったり，人々一般であることが示唆される場合があり，また，モダリティ表現中の用言について (無理に) ガ格を推定する必要がないこともわかる．

例えば，次の (1) では願望を表すモダリティ表現から「願う」のは筆者であることがわかる．(2) では当為を表すモダリティ表現から，その前の「できる」「付き合える」のガ格は人々一般であることが推測される．

(1) 国際人を多く輩出して ほしいと願わずにはいられない.

(2) コミュニケーションがきちんとでき なければならない．さらに，自然体で人と付き合えることが望ましい.

このように，相互に関係の深いモダリティ表現の扱いとガ格の推定という二つの問題を同時に扱うことが本稿の目的である．これらの問題では，取り扱う表現の網羅性が問題となり，一気に網羅性の高い高度な処理システムを実現することは難しい．そこで，まず基本的なシステムを構築して，その解析結果を人手で修正して正しい情報を付与したコーパスを作成し，同時にシステムを徐々に高度化するというアプローチをとる．

以下の節では，モダリティ表現検出とガ格推定のシステムについて述べ，またコーパス作成のインターフェースを紹介する．

意思：これからやろう，またはやってもらいたいと思う何かについての態度	
意志	～したい，～しよう
申し出	～させていただきます，～いたしましょう
勧誘	～しましょう，～しませんか
命令	～しなさい，～すること
禁止	～するな，～するんじゃない
依頼	～してください，～してくださいますか
願望	～して欲しい，～してもらいたい
当為：一般的常識に照らした事態の必要性，望ましさ	
当為	はずだ，べきだ，～ねばならない
	必要だ，望ましい，当然だ
	やむをえない，～ざるをえない
許可	～してよい，かまわない
判断：話者の判断の確からしさに関するもの	
推量	～だろう，～らしい
可能性	～かもしれない，～しそうだ
様態	～ようだ，～みたいだ
伝聞	～らしい，～という
その他（婉曲的表現など）	
	～と思う，～と考える， ～と見る，～とみられる， ～ことである，～ことなのだ， ～とされる，～だとされる， ～としている，～ことにしている， ～となりました，～となっている， ～といえる，～ということでもある， ～であろう，～にすぎない， ～だったからである，～がそうである， ～ことが目に見えている， ～ことができる

表 1: モダリティの分類と表現例

2 モダリティ表現の検出

いわゆる真性モダリティについては [3] に従い意志，当為，判断の 3 種類に大別する．その他の広い意味でのモダリティ表現については，分類を適切に行うことが難しいので現時点ではひとまとまりのカテゴリーとしている．これらのモダリティの分類，表現例を表 1 に示す．

これらの表現に対してモダリティのタグを付与する処理は，構文解析システム KNP[2] のモジュールを用いて行う．KNP には，言語表現とそれに与える feature(タグ) の一覧をルールとして与えれば，その言語表現を形態素解析・構文解析し，パターンマッチングの規則を自動生成する機能がある．この機能を利用すれば，ほぼ表 1 のような一覧を用意するだけで，入力文中のモダリティ表現部分にタグを付与することができる．現在，ルールとして列挙している表現数は約 600 である．

3 ガ格の推定

3.1 アルゴリズム

用言のガ格の推定について，以下のようなアルゴリズムのシステムを構築した (実際には KNP を拡張した)．

0. 文章を入力とする．順次，文の構文・格解析 ([4] の方法) を行い，文中の各用言に対して以下の処理を行う．
1. 格解析で，構文的に直接つながっている格要素の中にガ格と判断されたものがある場合は (その用言の) 処理を終了．格解析は，表層的に格助詞「が」を伴う場合だけでなく，未格 (提題助詞などで格助詞が欠落しているもの)，連体修飾先 (「宿題を出した先生」の「先生」) なども対象として，格フレーム辞書とのマッチングに基づきガ格を推定する．
2. 次のときは特別の処理を行う．
   - 意志，願望のモダリティのタグがついた用言については，筆者をガ格とする．
   - それ以外のモダリティのタグがついた用言については処理対象外とする．
   - 時間・状況などがガ格であると考えられる例外的表現についても処理対象外とする．

例)「昭和にはいって」「今日にいたり」
「元をただせば」「残念なことに」

3. 文内の探索
   文内の，対象用言に構文的に直接つながっていないガ格または未格の格要素を次の順で探索し，対象用言の格フレームのガ格制約をみたすものがみつかれば，それをガ格として処理を終了．
   (a) 並列関係にある節内
   (b) 構文木中の祖先の節を近いものから順に
   (c) 構文木中の子孫の節を近いものから順に
4. 文外の探索
   現在の処理対象文の前の文から順に，各文について文末から順に，対象用言の格フレームのガ格制約をみたすものがみつかれば，それをガ格として処理を終了．

ここで，格フレームのガ格制約とは，その用言の格フレームのガ格の用例と，対象名詞との間の類似度がある閾値を超えることとする．格フレームのガ格には複数の用例があるが，類似度はその中での最大値とする．類似度の計算にはNTT日本語語彙体系[5]を用い，語彙体系中の木構造内での2語間の距離を，最大値1に正規化したものとする．類似度の閾値は次節で示す実験では0.6とした．

3.2 実験

現在のシステムで，京大コーパス中の5社説について，それぞれの先頭20文，合計100文の解析実験を行った．解析では京大コーパスに与えられている(人手で修正された)構文構造の情報を利用した．すなわち，構文解析結果をシステムへの入力とし，その格解析からはじめて前節のアルゴリズムを適用した．この解析結果を表2に示す．正しい構文解析結果を入力としているので，用言にガ格の格要素が係っていればそれが正しいものであることは間違いない．また，表2に示すように，格解析による未格，連体修飾先に対するガ格推定は非常に高精度に行われている．これは利用している格フレーム辞書のよさを示すものであり，省略の推定においても格フレーム辞書のガ格制約を用いるのであるから，よい結果であるといえる．

本稿の中心的課題である，特別処理，文内，文外の探索については現在のシステムの解析精度はまだ5割程度である(解析の失敗例を下に示す)．我々のアルゴリズムの主眼はモダリティ等を考慮して行う特別処理

	ガ格	格解析 (未格・連体)	特別処理	文内	文外
正解	77	107	32	23	3
誤り	-	2	12	8	43

表2: ガ格推定の実験結果

の網羅性にあるが，これもまだ十分なものとはいえない．今後，解析結果の人手による修正を行いつつ，扱う表現の網羅性をあげていきたい．また，現在の文内，文外の探索のアルゴリズムは非常に単純であり，今後検討が必要である(これらについて詳細なアルゴリズムを考察したものとして[6][7]などがある)．しかし，この部分については，アルゴリズムを検討するというよりは，ある程度の学習コーパスを作成してから機械学習の手法を用いることが適当である可能性も高い[8][9]．

解析の失敗例

- 文中の探索の失敗
  (3)「坂の上に雲を見つめ，いまようやく坂の上に上り詰めた」と司馬遼太郎式に、戦後半世紀を経た日本経済の位相を解析した．

  下線部の用言のガ格は2つとも「日本経済」であるが，現在のアルゴリズムでは文中の(表層上の)ガ格，未格のみを探索対象としているために失敗する．

- 文外の探索の失敗
  (4) 昨年，一人の論客が世を去った．通産官僚時代「町人国家論」を説いた天谷氏である．

  今のアルゴリズムでは一文目の「世を」を二文目の「天谷氏である」のガ格と判定するが，正しくは一文目の「論客が」である．

4 解析結果の人手修正

これまでに示した方法によってモダリティ表現の検出とガ格の推定の自動処理を京大コーパスに対して行い(京大コーパスの構文情報はそのまま利用する)，その結果を人手で修正する．修正には，京大コーパス作成時のユーザインターフェースを拡張したものを用いる(図1)．

この人手修正作業は，相互に関係の深いモダリティとガ格の問題を同時に扱うので，作業者の集中力，興

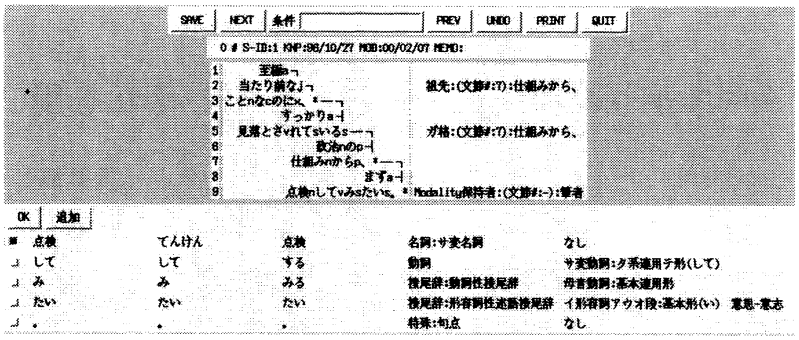

図 1: コーパス作成のユーザーインターフェース

味が適当に刺激され，一方の作業だけを行うことに比べて作業の精度が高まることを期待している．

また，この過程を通じて京大コーパスの誤りが見つかり，修正されることも期待している．これも，単に構文情報を見直すというタスクでは作業者に負担が大きいが，ガ格の省略などを調べながら文を読む作業の中であれば，自然に行えると思われる．

現在のところ，試験的な人手修正作業を通して，システムの初歩的な見落としの修正，インターフェースの改良などを行っている．来年度からは本格的な (日常的な) コーパス修正作業を開始する予定である．

5 おわりに

モダリティ表現の検出と，用言のガ格情報の推定をまず自動的に行い，その結果を人手修正していくプロジェクトについて述べた．これらの問題は，システムで扱う表現の網羅性が重要であり，そのためにはこのようなコーパス作成に沿ったシステム改良の方法が適していると考えられる．

モダリティと省略の扱いが実用レベルの精度で実現されれば，言語情報処理の可能性は大きく広がるものと期待できる．

参考文献

[1] 仁田義雄・益岡隆志 編 (1991). 日本語のモダリティ. くろしお出版.

[2] 黒橋禎夫, 長尾眞 (1994). "並列構造の検出に基づく長い日本語文の構文解析." 自然言語処理 Vol.1 No.1.

[3] 黒橋禎夫, 木下恭子, 山田悟史, 長尾眞 (1998). "文タイプと文間関係の情報を付与したテキストコーパスの作成." 言語処理学会 第 4 回年次大会発表論文集 pp.342-343.

[4] 河原大輔, 鍛冶伸裕, 黒橋禎夫 (2000). "大規模コーパスからの格フレーム辞書構築とそれを用いた格解析." 言語処理学会 第 6 回年次大会発表論文集.

[5] NTT コミュニケーション科学研究所 (1997). 日本語語彙大系. 岩波書店.

[6] 中岩浩己, 池原悟 (1996). "語用論的・意味論的制約を用いた日本語ゼロ代名詞の文内照応解析." 自然言語処理 Vol.3, No.4 pp.49-65.

[7] M.Murata(1996). "Anaphora Resolution in Japanese Sentences Using Surface Expressions and Examples" Doctoral thesis, Kyoto University.

[8] C.Aone, S.W.Bennett.(1995). "Evaluating Automated and Manual Acquision of Anaphora Resolution Strategies" In *33rd annual Meeting of the ACL* pp.122-129.

[9] K.Yamamoto, E.Sumita.(1999). "Multiple Decision-Tree Strategy for Error-Tolerant Ellipsis Resolution" In *5th Natural Language Processing Pacific Rim Symposium* pp.292-297.