大規模コーパスからの格助詞生成規則獲得実験

6G−1

野口　喜洋
(株)日本電子化辞書研究所（ＥＤＲ）

１．はじめに

意味ネットワークによる文意表現からの日本語文生成では、述語概念と他概念を結ぶ弧で表現される深層格関係から、表層の格助詞を決定する規則（以後「格助詞生成規則」と記す）が必要である。格助詞生成規則は、表層の述語の意味・用法毎に固有であり、深層格と表層格の対で示される格スロットと、フィラーとなる名詞概念の意味分類を必要に応じて記述した格フレームの形式を取る。このような規則は、生成対象の全動詞に対しあらかじめ高精度に記述するのは困難である。計算機支援のもとに実際の使用事例から半自動的に獲得し、知識処理システムでの運用を通じた評価データによって、インクリメンタルに改良していくのが実際的な方法論である。

本稿では、日本語文の形態素解析と意味解析結果の対を格納した大規模コーパスを用い、動詞を対象に格助詞生成規則を半自動的に獲得する実験を行う。結果をもとに使用事例収集や格助詞生成規則への統合における問題点について考察する。特に、高精度の規則を得るために、獲得過程において人間が判断を行うべき項目と、計算機で処理可能な項目の見極めを行うことを目的とする。

２．格助詞生成規則の獲得手順

事例は、日本語文を対象とし、文の形態素解析結果と人間による意味解析結果の対を格納したコーパス（ＥＤＲコーパス：データ例は図１）より収集した。形態素結果から表層の助詞と文法情報が、意味解析結果から、深層格と名詞概念の意味分類を事例データとして得る。

ANCがソ連から武器援助を受け、共産党員が加わっている
ことも西側諸国にANC援助に二の足を踏ませている。

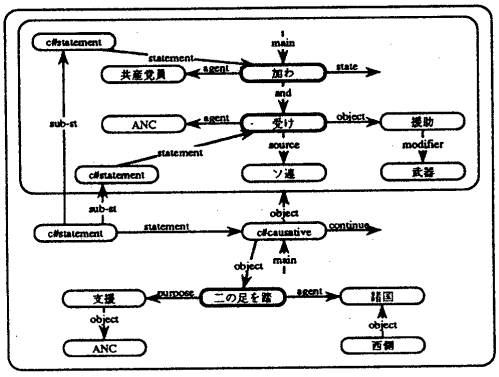

図１：コーパス中の格助詞情報と意味解析情報

本手法により獲得を試みる格助詞生成規則の形式を以下に示す。

〈格助詞生成規則〉::=
　〈単語見出し〉〈概念ＩＤ〉
　{〈深層格〉〈表層格〉{〈概念ＩＤ〉}}*

例）受け　0c59ca(ある働きかけに接して応じる)
　　agent が object を source から

格助詞生成規則は、動詞の意味・用法に対する格の共起パターンと、深層格／表層格の対応を記述する。

言語解析で曖昧性低減のために利用する場合と異なり、生成では、名詞概念の意味分類は既知であるため、ある意味・用法に複数の格助詞生成規則がある場合のみ意味分類を概念ＩＤとして記述し、規則選択の規準とする。

動詞によらない一般的な規則としての深層格・表層格の対応はあらかじめ格対応表として定義しておく。一つの意味・用法の規則は、この対応の部分集合となる。関係子は、必須格・任意格・格関係以外の関係に３分類され、格助詞生成規則には必須格の格パターンだけを記述する。格対応表は、事例から格助詞生成規則への統合の際、意味解析結果に起因するノイズを低減するため利用する。以下、獲得手順の各段階を説明する。

２．１　事例収集

表層文の格パターンは、以下の過程で決定されると仮定する。

格助詞生成規則のいくつかの
スロットを名詞概念が埋める

態(voice)による表層格の置換

副助詞等による格助詞の変形・置換

格助詞生成規則の獲得では、原始事例である表層文の格パターンから、上記の逆の過程をたどり、能動態で副助詞等を用いない格パターンを事例として収集する。

事例の形式は、格助詞生成規則と同じであるが、名詞概念の概念ＩＤは省略しない。

生成時には、格助詞生成規則により得た格パターンを元に、上記過程を制御して表層文の格パターンを決定する手法を用いる。

以下、事例収集の手順について説明する。

まず、コーパスより原始事例を収集し、動詞の概念ＩＤにより分類する。概念ＩＤは意味・用法を表しているため、多くの場合、格助詞生成規則は一つに統合できる。深層格・表層格・名詞概念の概念ＩＤの組を事例データとして収集する。動詞に後続する形態素を調べ、文の態・それに類する表現も事例データとして収集する。

つぎに、表層格の表現として、格助詞の前に副助詞が付与されている場合は削除する（例：だけに→に）。格助詞が副助詞で置き換えられている場合は、元の格助詞

An Experiment of Case Frame Acquisition from Large Scale Corpus
Yoshihiro, NOGUCHI
Japan Electronic Dictionary Research Institute, LTD.(EDR)

が推定できる場合を除き、格助詞不明のまま格の生起のみを事例データとする（例：も→？）。本手法では、終点を表す「まで」格を表層格の一つと考える立場を取るため、格対応表を利用して副助詞の「まで」を判定して処理している。

最後に、文の態により置換された表層格を可能な限り元に戻す。また、可能・希望・やりもらい動詞・「ようにいう」の voice 的側面を持つ表現も処理する。によっても変換される。態やこれらの表現は、原始事例収集時に判定されているので、以下の処理を行うことで、態の影響を排除する。使役受動態など、他の態と共起する場合も同様に処理する。

能動態・再帰態　置換しない。
受動態　受動態のタイプを判定し、置換する。
使役態　置換する。使役の指示者は使役を表す特別な概念（c#causative:図１参照）の agent として表現されるため、動詞の格としては記述しない。
相互態　相互態を取り得る動詞（例：結婚）は、日本語単語辞書中に表層格の「と」格を取ると記述されている。また、接尾辞「合う」によって派生した相互動詞（例：殴り合）の項目も存在するため、置換せず、格パターンの一つとして処理する。
可能・希望・やりもらい動詞・「ようにいう」　原始事例収集の際に可能な限り検出し事例としては収集しない。

２．２　格助詞生成規則への統合

収集した事例には、作業者による意味解析時のミスや判断のゆれに起因するノイズがある程度混在しており、格助詞生成規則への統合を行う際に悪影響を及ぼす。意味解析時のミスは、主に関係子選択と動詞の概念選択において観察された。

前者に対しては、格対応表の範囲外の深層格／表層格対応を事例データから削除することで対処した。

後者には有効な対策が少なく、ある程度の規模で事例が集まった時点で、極めて頻度の低い格パターンを削除するなどの処置が考えられる。今回は事例の数が少ないため、人間が例文と比較して判断した。

上記のノイズ対策を施した後、各事例をまとめて格助詞生成規則を獲得する。その手順は以下の通りである。

①事例Ａが事例Ｂの深層格／表層格の集合を全て含む場合、ＡをＢにまとめる。表層格が「不明」の場合は、任意の相手とまとめられる。
②フィラーである名詞概念の整合性を判断する。統合後は、Ａ・Ｂ両方の名詞概念の共通の上位概念が名詞概念欄に記述される。
③与えられた事例から、これ以上統合できないものを、動詞の格助詞生成規則とする。また、その時点における名詞概念を、意味分類とする。

３．実験と考察

実験に用いた事例の内訳を表１に、得られた事例からの格助詞生成規則の統合結果を表２に示す。

表１　事例の内訳

テキスト（文）の総数	事例数	動詞異なり数
1242	2294	1250

表２　格助詞生成規則の統合結果

項目＼対象	全1250動詞	事例数10以上の15動詞
事例数	2294 (1.84)	334 (22.3)
生成規則異なり数（１動詞あたり）	1869 (1.50)	123 (8.2)
統合後異なり数（１動詞あたり）	1487 (1.19)	63 (4.2)

意味・用法別に格助詞生成規則を統合したため、多くの場合１つの格助詞生成規則に統合できた。また、格助詞生成規則が２つ以上得られた場合の名詞概念の区別も、効率的につけることができた。態による助詞の置換も、問題なく機能している。

反面、本実験の問題点として、動詞あたりの事例数がまだ少なく、生起していない表層格があると思われるため、統合過程が途中で終わってしまった動詞もあること、必須格・任意格の区別が動詞の種類によらず固定であるため、ほとんどの動詞では任意格と思われる time-from, time-to, location, place などの深層格が統合を阻害した場合があることが挙げられる。

また、「ある」「なる」「いる」など、多くの意味・用法を持ち、さまざまな名詞概念と関係を持ちうる動詞は、本手法だけでは妥当な格助詞生成規則を得られないため、別に扱うことが望ましいことも判った。

名詞概念の統合については、概念間の上位－下位関係を格納した概念体系辞書によって自動的に行うのが望ましいが、本実験では、概念体系を参照して人間が名詞概念の意味的整合性と上位概念を判断した。名詞概念統合の自動化には以下のような課題があり、今後の実験で明かにしていく。

①概念の集合Ｓを包含する最も狭い（特殊化された）概念Ｃの妥当な定義。それと関連してノイズの事例をどのように扱うか。
②概念の集合Ａ・Ｂを一つの概念Ｃに統合すべきかどうかの判定規準の妥当な定義。

４．おわりに

日本語文の形態素解析と意味解析結果の対を格納した大規模コーパスから、日本語文生成に利用する格助詞生成規則を半自動的に獲得する手法を提案し、動詞を対象に実験を行って、有効性を確認した。

今後は、概念体系辞書を用いた名詞概念の意味分類の自動統合について実験を行い、本手法を完成させる。また、副助詞による表現の付与規則、態の決定規則等の獲得を行い、日本語文生成のためのコーパス利用技術の確立を目指す。

参考文献

〔1〕宇津呂，松本，長尾：二言語対訳コーパスからの動詞の格フレーム獲得，人工知能学会第６回全国大会論文集，15-6，pp.547 - 550(1992).
〔2〕小松：係り受け事例からの結合価フレーム獲得の試み，情報処理学会第４３回全国大会論文集，3H-2，pp.3-209 - 210(1991).
〔3〕高橋：現代日本語のヴォイスについて，日本語学，vol.4，pp.4 - 23(1985).
〔4〕野口：オブジェクト指向設計に基づく文生成システム，情報処理学会第４３回全国大会論文集，4G-5，pp.3-159 - 160(1991).