小学校教科書を対象とした日本語格解析システムの作成

奥村泰章　　白井英俊
中京大学 大学院 情報科学研究科　　中京大学 情報理工学部

１．はじめに

我々は、最終的に文章理解システムの構築を目標とし、その基盤となるシステムを試作した。このシステムは、形態素解析と構文解析、および格フレーム辞書を用いた格解析により、文が持つさまざまな情報を整理、記述するものである。本研究では、試作したシステムを用いて一般に解析の難度が低いと思われる小学校の教科書を解析し、その結果を分析することで、現状のシステムの解析可能な範囲と、解析できないパターンに対する解決策を考察するものである。それにより慣用表現などの特殊な用法を網羅した格フレーム辞書や、意味制約の変化に対応した意味体系とそれを利用する名詞の概念構造の辞書を構築する必要が確認された。

２．システム概要

形態素解析に mecab ver0.97、係り受け解析に cabocha ver0.60pre2[3]を用い、文の形態素情報や係り受け情報に基づき、日本語語彙大系[2]が持つ格フレーム辞書と入力文との照合による格解析を行う。出力は、入力文との格の対応が記述された格フレーム構造である。これにより省略が格フレームの中の未充足の格にとして認識される。

３．実験、評価

小学校の国語の教科書の説明文(3 年生下 1 本、6 年生下 2 本の計 232 文、3485 語)を対象に評価実験を行った。出力結果を人手で分析し、対象の述語に対して、文中で用いられている意味の格フレームが存在し、格の同定を正しく行えているものを正解とした。その結果、全体での正解率は 64%となった。

４．分析

解析に必要な情報や処理を明確にするために、失敗例を要因ごとに分類した。

・格助詞が現れない場合の誤認

連体修飾節や副助詞を伴う名詞句のように格マーカが現れない構文に対し、格の同定を誤認する場合があった。このような問題に対しては、基本的には格の意味制約を用いて判断するしかないが、「など」や「も」のような助詞は、代用となる格に制約があるため、可能な格の範囲の出現頻度などを用いることで解決できる。また、連体修飾節に関しては、それが関係節構造となっているかを認識しなければならないが、内容節や同格節の場合は、係先となる名詞の制約を利用して解決できると考えられる。

・辞書から格フレームが取得できないもの

「飲みほす」や「やって来る」などの複合動詞が辞書に登録されていない場合が３％あった。これは格フレーム辞書と形態素解析器の辞書との語の認定範囲が違っていることが原因であると考えられ、生成的な複合動詞の場合には、その構成要素まで分解できるよう統一されることが望ましい。

(1) これらの国々で構成されるアジアの面積は、世界の五分の一にもなり…(後略)
(6 年下)

(1)における「なる」は慣用的に「～である」と同じ意味で用いられている。このような述語の慣用句的な用法が９％あり、これらを扱うための辞書が必要である。

・意味制約の不整合

(2) この十万年間の地球上に起こったどんな変化よりも大きいのです。 (6 年下)

(2)の「起こる」は二格に具体物という制約を持つが、共起する「地球上に」は、日本語語彙大系の意味階層では「上」という抽象概念となり、意味制約の不整合が起こる。このような場合が４％あった。「上」は関係的な名詞であり、共起した語の素性を継承するものである。このような語を正しく認識し、意味素性を生成することが必要である。

(3) ブレーキをかけるときは、赤のランプがついて、後ろの車に知らせます (3 年下)

(3)の「知らせる」においては、「二格」の意味制約は主体であり、意味的には「後ろの車(の中の人)」である。このような語の概念的

な情報に基づくタイプの変更は、生成語彙論[1]で指摘されているように、体系的である。

・構文解析の失敗

cabochaによる構文解析が失敗しているために解析に失敗しているものが１％あった。

(4) 最も有効なのは、エネルギーを賢く利用する事、つまり、エネルギーを効率良く、効果的に利用するという事です。(6年下)

(4)の二番目の「エネルギーを」が「利用する」ではなく、「良く」に係るとする誤解析がおこった。このような「主動詞の格要素」＋「述語の連用形」＋「主動詞」というパターンに対しては、連用接続している述語の取りうる格関係を検証することによる修正が可能である。

５．省略要素の認識と処理

格の欠落を認識したならば、省略された要素が存在するかどうかを認識し、また存在するならばそれを補完する処理が必要である。そのために、解析結果において文中に格要素が現れない事例の分析を行った。

・動詞句の認定

(5) 効率を上げるというのは…(中略)……済ませるという事です。(6年下)

日本語は、表層上では動詞句と文の区別がつきにくく、例えば(5)の「効率を上げる」は主語が不要な動詞句である。このような表現は、その後に「～のは」や「～という」、「～ために」のような表現を伴うことが多く、そうしたパターンを認識することが必要である。

・任意格が文脈に存在している場合

(6) このまま大気中の二酸化炭素の濃度が増え続ければ、二十一世紀の前半には気温が地球全体で一度ほど上がり…(後略)(6年下)

(6)の「上がる」の「カラ」と「マデ」はそれぞれ範囲の始点と終点を表す。このような格は任意格であるが、(6)では「一度ほど上がる」という変化量によって始点と終点が含意される。また始点のみ、終点のみが現れる場合もあり、このような場合文脈から復元する処理が必要である。

・慣習的な省略

(7) 正月には、一年の始まりを祝い、その年の安全と幸福を祈る祭りや行事をします(4年下)

(8) 中国では、爆竹や花火を鳴らして、悪霊を追い払います。(4年下)

(7)の「行事をします」と (8)の「追い払います」の「ガ格」が省略されている。この時の「ガ格」は両方とも「一般の人々」であり、慣習的に省略されている。しかし、(8)の場合は「中国の一般の人々」であり、(7)の場合とは指示対象が異なっている。このような場合には、解析の際には文脈よりそれがどのような指示対象をもつかを判断し、区別することが必要である。

(9) これからは、「使い捨て文化」から「長持ち文化」に変わっていくべきだと思います。(6年下)

(9)では、「思います」のガ格は「筆者」が省略されている。また、「読者」などが省略されうる。このような省略は文章上で「思う」や「考える」などの述語、「呼びかけ」や「命令形」などの文体で判断できよう。

６．発展

社説などの大人向けの文章を解析するためには、現在の社会情勢などの特別な知識が必要となる。本システムを用いて子供向け新聞記事や百科事典などを解析することでそのような知識を自動で収集することができると考えている。また、述語とその格関係の情報は、文章構造のような事象間の関係を推測する処理が必要とする情報のうちの一つであり、本システムをベースとして述語間の関係を推定するようなシステムを考えていきたい。

謝辞

著者はこの研究のための資金援助をいただいた中京大学特定研究助成に感謝いたします。

参考文献

[1]Pustejovsky, J. (1995). *The Generative Lexicon.* Boston: The MIT Press.

[2] 池原ら (1997). 『日本語語彙大系 CD-ROM版』. 東京：岩波書店.

[3] 工藤拓, 松本裕治 (2002). チャンキングの段階適用による係り受け解析. 情報処理学会論文誌, 43(6). 1834-1842