コスト付き単一化を用いた文法的不適格文処理

今一 修　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

文法的不適格文を処理するための様々な手法が提案されているが、それらは、統語的な情報だけを利用したものや、意味的な情報だけを利用したものなど一部の不適格性を対象にしたものがほとんどであった。より柔軟に文法的不適格文を処理するためには、統語情報、意味情報、文脈情報を統合的に利用することが必要である。本稿で提案する手法では、単一化文法の枠組みを基本とし、統語、意味、文脈情報を素性構造を用いて統一的に扱う。

不適格文処理は部分解析結果を選択することによって行なう。この選択を統一的に行なうために、**コスト**と**報酬**という概念を導入する。コストと報酬という概念は関連性理論 [3] で用いられており、本稿でもそれと類似した考え方を用いる。コストはその部分解析結果の悪さ、報酬は良さを示している。部分解析結果はコスト最小・報酬最大の原則に従いもっとも適切なものが選択される。コストは、本稿で導入する**コスト付き単一化**により計算される。本稿では、不適格文処理を行なう3つのモジュールについて述べる。**モジュール A** は、制約違反などの局所的な不適格性を扱う。**モジュール B**は、語順の間違いなど非局所的な不適格性を扱う。**モジュール C**は、上の2つのモジュールからの結果を入力とし、省略などの文脈情報を必要とする不適格性を扱う。以上の3つのモジュールがコストと報酬に基づく統一的な枠組みで動作することについても述べる。

2 不適格文のタイプ

文法的不適格文が実際のコーパスにどのように出現するかを調べるために、ATR 対話データベース [7] を分析した。分析対象は、国際会議の申し込みに関する参加者と事務局の対話 (会議タスク) における電話対話文1000文とキーボード対話文1000文である。分析結果を表1に示す。以下、簡単にそれぞれの言語現象について述べる。

語句の欠落　これは必要な語句が欠落している場合で、助詞の欠落と格要素の欠落がある。本分析では、電話対話文の61.9%、キーボード対話文の51.4%の文で語の欠落が発生している。日本語の会話文では、話者、聴者が省略される傾向があるが本分析においてもこれを裏付ける結果が得られた。また、助詞欠落において、どの助詞が欠落されるかを分析した結果、必須格要素の助詞が欠落する場合がほとんど (電話データで約97%、キーボードデータで約89%) であった。

表 1: コーパスに出現する不適格文の分析結果

不適格な現象	会議タスク	
	電話	キーボード
語句の欠落	619	514
助詞	113	26
格要素	561	495
話者	445	328
聴者	170	185
主題	39	51
余分な語句	644	4
挿入句	10	1
間投詞	635	3
自己修復	256	0
制約違反	9	2
語順の誤り		
倒置	5	0
省略	21	69
少なくとも一つ	810	581
計	1000	1000

余分な語句　音声対話文においては、「あー」、「えーと」などの間投詞や自己修復文 (言い直し) が頻繁に出現する。自己修復とは、「私は京都へ、えー、京都大学へ行きます」のように文の途中で発話を一旦中断して語句を言い直すことである。本分析でも、これらの現象が音声対話文に特徴的な現象であるという結果が得られた。

制約違反　日本語には統語的制約が少ないため、制約違反はほとんど出現していない。意味的制約違反としては、比喩、換喩、擬人化などがある。

語順の誤り　語順の誤りの典型例は倒置である。日本語では、動詞が下位範疇化する名詞句[1] の語順は比較的自由である。例えば、「太郎が学校へ行く」も「学校へ太郎が行く」も文法的である。しかし、動詞が下位範疇化する名詞句とその動詞の語順は自由ではない。つまり、それらの名詞句は動詞の前に出現しなければならない。

[1] 本稿で用いる文法の枠組みでは、名詞+助詞 (後置詞) を後置詞句とは考えずに名詞句と考えている。つまり助詞が主辞ではなく、名詞が主辞である。

例えば、「行く、学校へ太郎が」や「太郎が行く、学校へ」などは非文法的である。本分析では、倒置は0.5%しか出現しておらず、頻出する言語現象ではないことがわかる。

省略 ここで言う省略は、談話においてそれ自体では不完全で断片的な文法単位 (主として名詞句) を指している。つまり、発話全体の統語カテゴリが文になっていないという点で、ゼロ代名詞化のような格要素の欠落とは区別している。

以上の分析をもとに本稿では不適格文を以下のように分類する。

タイプ 1 制約違反
タイプ 2 構造違反
タイプ 3 文脈情報を必要とする不完全で断片的な句

タイプ 1 は文法が課しているさまざまな制約条件 (統語的制約、意味的制約) の違反である。助詞欠落もタイプ 1 に属するものと考える。本稿で用いる枠組みでは、動詞の下位範疇化要素である名詞句は助詞でマーキングされていなければならないとしている。従って、助詞が欠落した名詞句はこの制約条件に違反していることになるため、助詞欠落をタイプ 1 に分類する。日本語では制約自体が少ないため、このタイプに属する言語現象の数は少ないが、英語では、種々の統語的制約があるため、このタイプに属する言語現象の数は多い。

タイプ 2 は文法が課している制約条件をすべて取り去っても処理できないものである。つまり、その文法では許されていない品詞の並びに相当する。これには、語順の間違い (倒置文) が分類される。

タイプ 3 は一文内では処理できない言語現象である。これには、省略、格要素の欠落 (ゼロ代名詞化) が含まれる。

本稿では、以上 3 つのタイプを扱い、自己修復、間投詞などの余分な語句は扱わない。

3 基本アイデア

前節で述べた不適格文に対処するためにこれまでにいくつかの手法が提案されてきた [6]。不適格文を扱う方法としては、1) 文法を拡張する方法、2) 文法は拡張せずに別のメカニズムを用いる方法、が考えられる。本稿で採用するのは 2) の方法である。

2) の方法の例として**緩和法**があげられる [4]。緩和法では、制約違反により文の解析に失敗した場合、違反した制約条件を緩和することによって処理を行なっている。例えば、"**John love the girl*" のように主語と動詞の一致に関する制約を違反している場合を考えてみる。緩和法の枠組みでは、まず通常の処理が行なわれる。しかし、この文は主語と動詞の一致に関する制約に違反しているために解析に失敗する。このような場合に緩和法では、違反した制約を緩和して処理をもう一度行なっている。

この緩和法には以下の欠点がある。1) 一度適用して失敗した文法規則を制約を緩和してもう一度適用している。2) 違反している制約を緩和した結果、何が得られるのかが事前にはわからない。3) どの制約を緩和すべきかがわからない。

このような欠点を克服するために、制約違反によって文法規則が適用できない場合に、単に失敗に終わらせるのではなく、不適格性を含んだ構造をつくることを考える。この不適格性を含んだ構造には、a) どの程度、不適格か、b) 何が不適格か、という情報を持たせる。もちろん、この不適格性を含んだ構造は通常の解析には利用しない。後の不適格文処理のために保存しておくのである。

本稿で提案する枠組みは、HPSG[2] に基づく日本語文法およびチャート法 [1] を採用している。チャート法を用いることにより部分解析結果を保持することができ、不適格文処理の際にそれを利用することができる。また、HPSG に基づいた日本語文法を用いることにより、統語情報、意味情報を統一的に扱うことができる。

4 節では、この不適格性を含んだ構造をつくるための**コスト付き単一化**について述べ、5 節で不適格性を含んだ構造をどのように用いるかについて述べる。

4 コスト付き単一化

単一化文法の枠組みでは、単一化の成功/失敗によって制約の充足/違反を捉えている。古典的な単一化は、単一化を行なう素性構造間に矛盾した情報が検出されると単一化に失敗する。つまり、古典的な単一化は矛盾した情報を扱うことができない。本稿では、不適格性を含んだ構造をつくるために、古典的な単一化を拡張し、矛盾を含んだ情報を扱えるようにする。この拡張した単一化を**コスト付き単一化**と呼ぶ。

コスト付き単一化は、素性構造間に矛盾した情報が検出されない限り古典的な単一化と等価である。素性構造間に矛盾する情報が検出された場合は、古典的な単一化と異なり、コスト付き単一化は演算に成功する。コスト付き単一化の結果として得られる素性構造にはその矛盾の度合に応じて**コスト**が割当てられる。また、どこに矛盾が生じているかを**矛盾集合**を用いて表現する。

例えば、素性構造 (1) と素性構造 (2) のコスト付き単一化について考えてみる。

$$\begin{bmatrix} \text{NUMBER} : \textbf{singular} \\ \text{PERSON} : \textbf{third} \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} \text{NUMBER} : \textbf{plural} \\ \text{PERSON} : \textbf{third} \end{bmatrix} \quad (2)$$

この 2 つの素性構造は、NUMBER 素性に矛盾が生じて

いるために古典的な単一化では演算を行なうことができない。しかし、コスト付き単一化を用いると、演算の結果として素性構造 (3) を得ることができる。

$$\begin{bmatrix} \text{NUMBER} : \top\{\textbf{singular}, \textbf{plural}\} \\ \text{PERSON} : \textbf{third} \end{bmatrix} \quad (3)$$

$\top\{\textbf{singular}, \textbf{plural}\}$ は矛盾集合と呼び、どのような矛盾がどこに生じているかを表現している。記号$\top$は矛盾を表している。

コストの割当てに関しては、各素性ごとに重みをつけるなどの方法が考えられるが、本稿では、矛盾が生じた素性の数をそのままコストとして計算している。従って、素性構造 (3) のコストは 1 になる。

5 部分解析結果の選択

コスト付き単一化を用いることにより、解析途中でコスト付きの部分解析結果がつくられる。入力文が文法的な場合は、コスト付きの部分解析結果は利用しないが、入力文が不適格な場合は、これらを利用して不適格文処理を行なう。基本的な処理方法は、コスト付き部分解析結果の中から解析の失敗を回復するために適切なものを選択することである。これは、緩和法で言えば、コスト付き部分解析結果の中で生じている制約違反を緩和することに相当する。以下では、どのようにして適切なコスト付き部分解析結果を選択するかについて述べる。

モジュール $\mathcal{A}$ モジュール $\mathcal{A}$ では、タイプ 1 の不適格性、つまり助詞欠落および制約違反を扱う。解析の失敗を回復するために適切な部分解析結果を選択するために以下を基準を用いる。

コスト最小・報酬最大の原則

コスト最小・報酬最大の部分解析結果を選択せよ。

報酬は以下の基準によって定義される。

- 報酬の基準 A
  1) 動詞句
  2) 名詞句
  3) その他の句
- 報酬の基準 B
  1) 共通の主辞をもつ句のうちで入力文の最も広い範囲を覆う句
  2) 最も右側にある句

この報酬の基準は、高い報酬が割当てられる順序を示している。報酬の基準 A だけでは最大の報酬をもつものが決定できない場合は、報酬の基準 B が用いられる。例えば、同じ動詞句でもより広い範囲を覆う句に高い報酬が割当てられる。

モジュール $\mathcal{A}$ の動作をみるために「私は、その本読みました」という文について考えてみる。この文は名詞句「その本」に助詞が欠落しているために解析に失敗する。コスト付き単一化により部分解析結果 A として動詞句「その本読みました」がコスト c_1、部分解析結果 B として動詞句「その本読みました」がコスト c_2で生成される。この 2 つの部分解析結果の違いは、A では、名詞句「その本」が動詞句「読みました」の目的語の役割にあり、B では、名詞句「その本」が動詞句「読みました」の主語の役割にあることである。意味的な制約により名詞句「その本」が動詞句「読みました」の主語の役割を占めるのは不適切である。従って、部分解析結果 A と B のコストの大小関係は $c_1 < c_2$となる。モジュール $\mathcal{A}$ は、コスト最小・報酬最大の原則に従い、部分解析結果 A を選択する。

モジュール $\mathcal{B}$ モジュール $\mathcal{B}$ では、タイプ 2 の不適格性、つまり構造違反を扱う。HPSG の枠組みでは、統語情報として SUBCAT 素性にその句が統語的に完成するためにどのような句が必要か、MOD 素性にその修飾句がどのような句を修飾するかが記述されている。また、意味情報として CONTENT 素性にその句が意味的に完成するために必要な要素は何かが記述されている。モジュール $\mathcal{B}$ では、選択された部分解析結果がもつこのような統語情報、意味情報を用いて処理を行なう。

モジュール $\mathcal{B}$ が起動されると、次の基準に従ってその句を完成させるために必要な句の探索を行なう。

句の探索基準

選択された句が要求している句を探索せよ。

例えば、「私は読みました、その本を」という文を考えてみる。動詞句「私は読みました」と名詞句「その本」を結びつけようとする文法規則はないため、この文はモジュール $\mathcal{A}$ では扱うことができない。コスト最小・報酬最大の原則に従えば、部分解析結果として動詞句「私は読みました」が選択される。この動詞句の SUBCAT 素性をみることにより、この句が完成した句になるためには目的語の名詞句が必要であることがわかり、句の探索基準により名詞句を探索する。この探索は非局所的に行なわれる。探索により名詞句「その本を」が発見され、動詞句「私は読みました」が目的語に課す種々の制約条件を調べ、それらがすべて満足されると動詞句「私は読みました、その本を」が生成される。この例において名詞句が「その本」である場合、動詞句「私は読みました」が目的語に課す制約条件を満足しないためにコスト付きの動詞句「私は読みました、その本」が生成され、モジュール $\mathcal{A}$ が起動されることになる。

アルゴリズム 以上のモジュール $\mathcal{A}$、モジュール $\mathcal{B}$を行なうための全体の構成を図 1に示す。図において矢印は部分解析結果の流れを示している。全体は 3 つの部分から構成されている。統語解析部、メタプロセス、不適格文処理プロセスである。

統語解析部では、統語的な処理だけを行ないコスト付き単一化は実行しない。メタプロセスは、統語解析部か

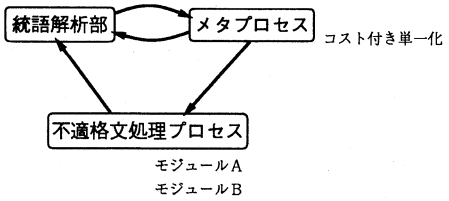

図 1: 全体の構成

ら部分解析結果を受け取りコスト付き単一化を実行する。コスト付き単一化により新たにコストが加算された場合は部分解析結果を統語解析部には戻さない。コストが加算されなかった場合には統語解析部に部分解析結果が渡される。メタプロセスから統語解析部に渡された部分解析結果をメタプロセスに**認可されている**部分解析結果と呼ぶ。渡されなかった部分解析結果を認可されていない部分解析結果と呼ぶ。つまり認可されている部分解析結果は統語解析部で用いられる部分解析結果である。

何らかの不適格性が原因で文の解析に失敗した場合には、不適格文処理プロセスが起動される。不適格文処理プロセスでは、部分解析結果の中からコスト最小・報酬最大の原則に従って不適格文の失敗を回復するために適切な部分解析結果を選択する。選択した部分解析結果がメタプロセスに認可されている場合は、モジュール $\mathcal{B}$ が起動され、認可されていない場合はモジュール $\mathcal{A}$ が起動される。モジュール $\mathcal{A}$、モジュール $\mathcal{B}$ において本節で述べた処理が行なわれ、その結果の部分解析結果が再び統語解析部に渡され処理が続行される。

6 文脈情報の利用

モジュール $\mathcal{A}$、モジュール $\mathcal{B}$ は一文内の不適格性を扱い、文全体を覆う句を構成するための機構である。モジュール $\mathcal{C}$ はその結果を受け取り、文脈情報を用いてその文の適切な解釈を行なう。例えば、「誰が来たの？」「太郎が」という省略の生じている発話対を考えてみる。通常の解析により「太郎が」が文全体を覆う名詞句であることがわかる。したがって、モジュール $\mathcal{A}$、モジュール $\mathcal{B}$は作用しない。このような断片的な発話に関して省略要素 (ここでは動詞「来た」) を補うのではなく、断片的な発話がその文脈上で適切であるとみなせればよいと考える。この例では「太郎が」という名詞句の断片が「誰が来たの？」という質問に対する適切な答えであればよい。「太郎が」は「誰が来たの？」の意味表現において具体化されていない動作主の役割を埋めるものとして統語的、意味的に適切であるので、この場合「太郎が」という名詞句の断片はこの文脈において適格であるとみなすことができる。

7 おわりに

本稿では、文法的不適格文を統合的に処理する枠組みを提案した。また以上の枠組みのうち、モジュール $\mathcal{A}$ とモジュール $\mathcal{B}$ については、計算機上への実装も行なった。このモジュール $\mathcal{A}$、モジュール $\mathcal{B}$ を用いることにより 2 節で述べた不適格性のうち、助詞欠落、制約違反、倒置を扱うことができる。日本語では制約違反がそれほど生じないが、英語では制約違反が頻繁に生じる。モジュール $\mathcal{A}$ は、このような制約違反に対して非常に有効な手法である。モジュール $\mathcal{B}$ も、制約違反に起因しない不適格性を扱うために有効である。それぞれ適用される言語現象が異なる 2 つのモジュールが、本枠組みでは、コスト最小・報酬最大の原則に従って統合的に制御されている。

モジュール $\mathcal{C}$ では、文脈情報を必要とする格要素の欠落、省略を扱う。しかし、現在のところ、まだモジュール $\mathcal{C}$ は完全ではない。今後は、どのように文脈情報を保持し利用していくかについても熟考する必要がある。これについては、平沢ら [5] が提案している関連性理論を用いた方法を利用する。

また本稿で取り扱わなかった言語現象として間投詞、自己修復がある。これらの現象は、音声対話文の処理を考える際には見過ごすことのできないものである。今後は自己修復や間投詞を扱うモジュールを考え、本稿で提案した枠組みに取り入れていく予定である。

参考文献

[1] Kay, M. Algorithm Schemata and Data Structure in Syntactic Processing, Technical Report CSL-80-12, Xerox PARC (1980).

[2] Pollard, C. and Sag, I. A. *Head-Driven Phrase Structure Grammar*, The University of Chicago Press (1994).

[3] Sperber, D. and Wilson, D. *Relevance - Communication and Cognition*, Blackwell (1986).

[4] Weischedel, R. M. and Sondheimer, N. K. Meta-rules as a Basis for Processing Ill-Formed Inputs, *Computational Linguistics*, Vol. 9, No. 3-4 (1983), pp. 161–177.

[5] 平沢純一, 松本裕治 関連性理論を用いた発話の解釈, 情報処理学会第 50 回全国大会 (平成 7 年度前期) (1995).

[6] 松本裕治, 今一修 頑健な自然言語処理の研究動向と課題, 情報処理学会第 1 回音声言語情報処理研究会 (1994).

[7] 江原暉将, 井ノ上直己, 幸山秀雄, 長谷川敏郎, 庄山富美, 森元逞 ATR 対話データベースの内容, Technical Report TR-I-0186, ATR 自動翻訳電話研究所 (1990).