文脈情報を利用した不適格文の構文解析

2R－2

那須川哲哉
日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

一般的な自然言語処理システムにおいては，構文解析の結果が処理のベースになることが多く，正しい解析結果を得ることが，システムの精度向上の要となっている．ところが人間が扱う全ての文を完全に解析できるような文法は存在せず，頑健な自然言語処理システムを構築する上では，文法的な誤りや省略を含んだ文，また非常に長い文など，予めシステムに記述した文法にそぐわない文(不適格) にも対処する必要がある．

本稿では，文法知識に依存した従来の構文解析処理では解析できない不適格文を，同じ文脈中で構文解析できた文 (適格文) の解析結果を利用して再解析する手法を提案する．また本手法を機械翻訳システム上に実現し，技術文書を翻訳した結果を通じて，その有効性を示す．

2. 不適格文の構文解析

不適格文の解析に関しては様々な試みが行なわれ，例えば，文法的な誤りに対処するため，数や性の一致といった文法的制約を緩和させて再解析する手法や，断片的な解析結果の組合せを出力する手法等が存在する [5]．しかしながら，そのような手法は，基本的に何らかのヒューリスティックスを用いて，一般的に可能性が高いと解釈される構造を構築するにすぎない．したがって，品詞や係り先選択の根拠が弱いため，解析結果の精度が低い．

一方，技術文書など，特定の対象に関する説明的な文章においては，内容に依存した語の用い方や筆者の語彙に，ある程度の一貫性が見られ，同じ文脈中では，同じ単語列が繰り返し出現し，その単語列の単語は同じような依存関係を結ぶという傾向が見られる [2, 3]．そのため，不適格文において，その構成要素の依存関係を解析する上で，同じ文脈中の他の文にその構成要素と同じ単語列が存在し，しかもその文が適格文として解析されている場合には，その解析結果から単語間の依存関係を抽出し，不適格文の解析に利用しようというのが，本手法の基本的な考え方である．具体的には，適格文の解析結果を文脈情報として蓄積しておき，不適格文中の部分的な単語列の係り受け関係を文脈情報から抽出することで，不適格文中の各単語が文脈に即した係り受け関係を結ぶように構文解析を行なう．

3. アルゴリズム

処理は大きく２段階に分かれる．第１段階では，文脈中の全文を構文解析し，解析に成功した適格文の情報を文脈情報として蓄えると共に，解析に失敗した不適格文を抽出する．不適格文に関しては，ボトムアップ・パーザを用いることで，部分的な解析構造を残すことができる．この部分的な解析構造は，パーザの文法では文全体として一つにまとまらないものの，部分構造自体はパーザの文法にかなっており，少なくとも部分的には正しい可能性が高い．したがって第２段階では，不適格文の構造を最初から解析し直すのではなく，この部分構造を修正しながら結合させて文全体の解析構造を構成する．こうすることにより，パーザの持つ文法を尊重すると同時に処理効率を向上させることができる．

第２段階では，第１段階で抽出した各不適格文の部分構造を，文脈情報を参照しつつ再構成する．ここではまず，各部分構造内での多品詞語の品詞選択や単語間の係り受け関係を文脈情報と照合し，異なる場合には文脈情報に即して修正する．次に，部分構造どうしが，文脈情報に存在する係り受け関係で接続できないかを調べ，可能な場合は接続する．この接続可能性を調べる際には，まず全く同じ語どうしでの接続パタンを検索し，それがない場合は，類義語での接続パタンを調べ，それもなければ，同じ品詞での接続パタンを調べるという形で，次第に制約を弱めながら各部分構造を結合していき，なるべく全体がひとつの構造にまとめられるように処理を行なう．

4. 実験結果

本手法を英日機械翻訳システム Shalt2[1] 上に実現し，技術文書を対象に実験を行なった．まず，実際に文脈情報を利用することで，不適格文から正しい解析結果を得られた例を示す．以下の入力文をＥＳＧパーザ [4] により解析したところ，図１のような二つの部分的な依存構造が得られた．

Fig. 3 is an isometric view of the magazine taken from the operator's side with one cartridge shown in an unprocessed position and two cartridges shown in a processed position.

図１の各々の解析構造において，文脈情報と比較して特異な解析がないかどうかを調べると，多品詞の選択に関しては side が該当し，

Completing Partial Parses of Ill-formed Sentences on the Basis of Discourse Information
Tetsuya NASUKAWA
IBM Research, Tokyo Research Laboratory

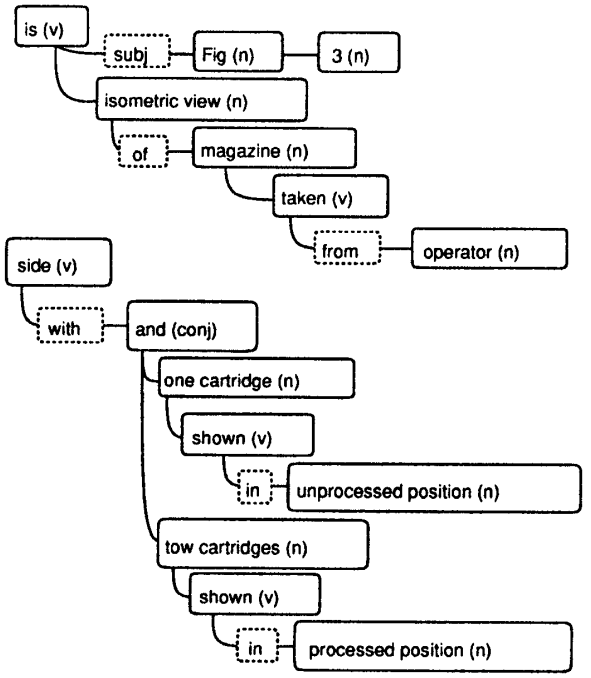

図 1: ESG パーザによる不適格文の解析出力例

- 適格文として解析された文中で, side は 15 回出現し, その全てが名詞として解析されている
- 名詞の side に対して cartridge が with を介して接続するパタンが存在する

という情報から, side は名詞に変更される. さらに係り受けに関しては,

- operator's が side に係る接続パタンが存在し, operator が from を介して take に係るパタンは存在しない
- side が from を介して take に 係る接続パタンが存在する

という情報から, take from operator の係り受けを外し, operator's を side にかける. この係り受けにより, 図 2 のように構造が一つにまとまるので処理を終了する. 図 2の解析構造からは正しい翻訳結果を得ることが出来た.

2 種類の文書 (特許文書及びマニュアル文) における不適格文を構文解析し機械翻訳した結果を表 1に示す. 本手法は, 文脈中で同じ語が繰り返し用いられ, その際に同じ語義を取り, 同じような語と係り受けを結ぶという性質に依存しているため, どの様な文章においても有効なわけではないが, 表 1に見られるように, 技術文書においては良好な結果が得られた. しかも, 何度も繰り返し出現する語句は文章中で重要な役割を果たしている場合が多く, 特にそのような語に関する解析精度を向上させられるという点で, 本手法の実効性は高い.

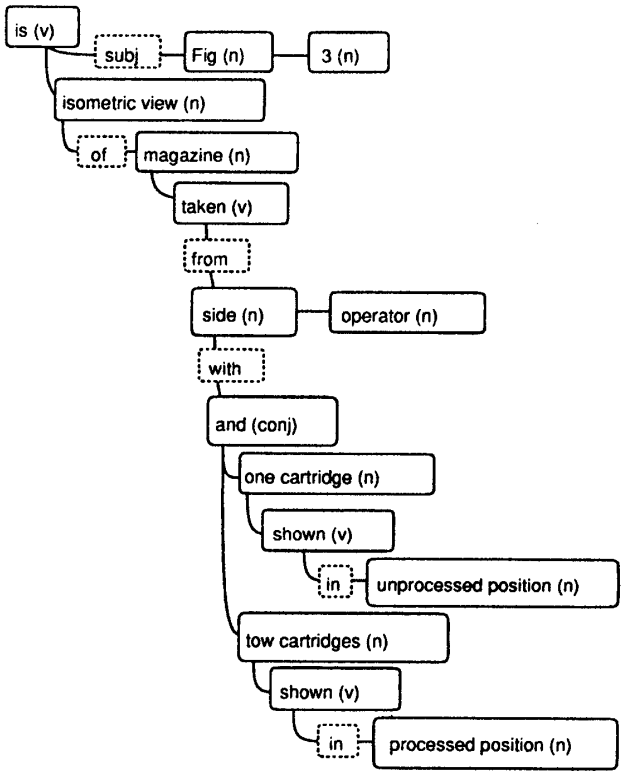

図 2: 文脈情報を利用した不適格文の解析出力例

参考文献

[1] Takeda, K., Uramoto, N., Nasukawa, T., and Tsutsumi, T. (1992). Shalt2 – A Symmetric Machine Translation System with Conceptual Transfer. *In Proceedings of COLING-92.*

[2] Nasukawa, T. (1993). Discourse Constraint in Computer Manuals. *In Proceedings of TMI-93.*

[3] 那須川哲哉 (1993)「文脈制約と文脈選好を利用した文脈処理システムＤＩＡＮＡ」, 情報処理学会自然言語処理研究会 NL98-8

[4] McCord, M. (1991). The Slot Grammar System. *IBM Research Report*, RC17313.

[5] 松本裕治 (1993)「頑健な自然言語処理の現状と動向」, 第 7 回人工知能学会全国大会

表 1: 文脈情報参照による不適格文の解析結果

		文書 1	文書 2
文脈中の総文数		175	354
不適格文の数		32	31
一つの構造になった文		18 (56.3%)	17 (54.8%)
翻訳文の品質の変化	向上	7	7
	同等	10	7
	悪化	1	3
部分的には変化した文		12 (37.5%)	8 (25.8%)
翻訳文の品質の変化	向上	4	2
	同等	7	3
	悪化	1	3
変化しない文		2 (6.3%)	6 (19.4%)