多義多品詞選択ルールを採用した依存構造解析

2R－3

青沢 秀憲　石井 利幸　笹野 明子　高木 朗
（株）ＣＳＫ

１．はじめに

入力文が長かったり多義・多品詞語を多数含む場合、探索空間が膨大になるため、単純なアルゴリズムで構文解析すると多大な実行時間を要し、又、解の候補も多くなるため正解も得られ難い。それ故、これまでにも高速、高精度化手法が数多く提案されてきたが、句構造解析をベースにした手法がほとんどである。

我々は、単純な依存構造解析アルゴリズムを採用した上で、依存関係を判定する為の従来からのルールとは別に、別の解の可能性を判定し選択する機能を持つ「多義多品詞選択ルール」を新たに導入し、探索空間を圧縮して高速、高精度化を図る手法を検討している。本稿では「多義多品詞選択ルール」を中心に、我々の解析手法やルール内容、そして解析実行結果について述べる。

２．解析手法の概要

我々は、バックトラックのある縦型探索方式で依存構造解析を行い、文頭から順に隣合う２つの部分木(語)に着目して依存関係を判定していき、文頭から文末までの語が１つにまとまった最初の解析木を正解として出力するという、単純な解析アルゴリズムを用いている。

しかし、長文や多義・多品詞語を多く含む文をそのまま素直に実行すると、バックトラック量が膨大になる。また、依存関係の判定のための単純なルールだけでは、着目する部分木以外の情報を参照するのが原則的に困難であるため、別の多義・多品詞語の組み合わせによる解候補との比較は行なえない。従って、出力する解析木も正解となりにくい。

そこで依存関係を判定するための従来からのルールとは別に、着目する部分木(語)以外に、隣接する語や多義・多品詞語の情報も参照することで、別の解の存在の可能性やその優先関係を判定する機能を持った「多義多品詞選択ルール」を導入した。これにより、単純な解析アルゴリズムを維持したままで、探索空間を大幅に低減でき、併せて、より高精度な解析が可能となった。

３．多義多品詞選択ルール

３．１　ルールの判定内容

本手法における依存関係判定ルールは、着目する２つの部分木が渡され、それらの間に依存関係が成立するかどうかを判定し、成立するなら「係る」(及びその依存関係の種別)を、成立しないなら「係らない」を返す。

これに対して多義多品詞選択ルールは、着目する２つの部分木と依存関係判定ルールの判定結果が渡され、

(1)別の多義・多品詞語の組み合わせを優先
(2)文法的にあり得ない並びの組み合わせを棄却

するかどうかを判定する。そして(1)の場合には、着目している部分木に含まれる特定の語を、指定した多義・多品詞語に交換する旨の「指定交換」(及び交換対象)を返し、(2)の場合には着目している部分木に含まれる最も文末側の語から優先的に交換する旨の「交換」を返す。尚、(1)、(2)に該当しない場合は、依存関係判定ルールの判定結果をそのまま返す。図１に解析アルゴリズムの概要を示す。

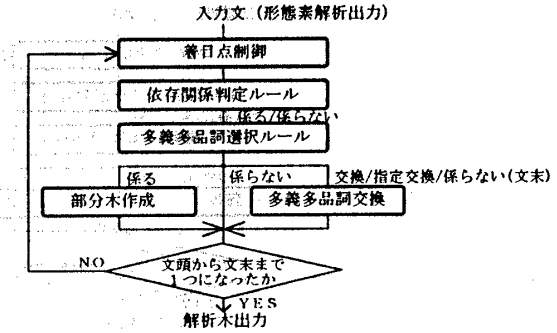

図１：解析アルゴリズムの概要

３．２　ルールの参照内容

多義多品詞選択ルールは、着目する２つの部分木のトップの語の品詞の組み合わせに対応するものが起動される。ルール内では、依存関係判定ルールの判定結果や着目する部分木ばかりではなく、文頭側に隣り合う部分木や文末側に隣り合う語の情報も参照できる。また、それらの部分木のトップの語以外に、各部分木に係っている語の情報も参照できる。しかも、それらの部分木に含まれる全ての語と並立して存在する多義・多品詞語情報も参照できるようになっている。以上のように、参照内容に物理的な制限は特に設けていないが、実際に参照するのは主に以下のような限られたものだけである。

(1)依存関係判定ルールの判定結果
(2)トップの語の品詞や活用形
(3)トップの語と並立する多義・多品詞語の情報
(4)着目する部分木に係っている語の有無、係っている場合は、その依存関係（格や属性関係）の種別
(5)読点の有無
(6)トップの語がとれる格や属性の制約（依存関係の種別、係れる語の意味素性、格マーカ、etc.）

少し複雑な（個別的な）ルールになると、まれに

(7)部分木に含まれる語の意味素性や見出し
(8)着目する部分木内の語と、隣接する部分木内の語との依存関係あるいは共起関係

等を参照することもあるが、殆どの場合は(1)～(6)の情報だけであり、それだけでも高速、高精度化を図ることにおいて十分な効果が得られている。

３．３　ルールの例

ルール例(概略)を表１に示す。便宜上、着目している部分木のうち、文頭側の部分木(表中〔〕)のトップの語を「前」、文末側の部分木(表中｛｝)のトップの語を「後」と表現し、又、「後」の文末側に隣接している語を「次」(表中〈〉)と表現する。表１は多品詞語選択の例であるが多義選択の場合も意味素性や、格、属性などの制約を参照して全く同様に行われる。尚、現在の多義多品詞選択ルール数は116、依存関係判定ルール数は53である。但し、これらのルールは異なる品詞の組み合わせ

Dependency Structure Analysis with Ambiquity Reduction Rules
Hidenori AOSAWA, Toshiyuki ISHII, Akiko SASANO, Akira TAKAGI
CSK Corporation

において共有されるものが多く、実質的には前者で1466、後者で440相当に値する。

表１：多義多品詞選択ルールの例

（A）動詞を棄却して助詞を選択（名詞ＶＳ動詞ルール）
【概略】前の名詞が後の動詞（連用中止）に係らない場合、前と後の間に読点がなく、後の多品詞語に副助詞があり、後に何も係っておらず、次に読点か動詞か格助詞があるなら後を副助詞に交換する。但し、次に助動詞がある場合は曖昧なので交換しない。
【適用例文】
〔この問題〕〔のみ〕〈議論する〉。
〔新聞〕〔さえ〕〈読ま〉ない。
〔教科書〕〔のみ〕〈で〉勉強する。
【非適用例文】
三井販売下げ、〔ダイア建、洋製鋼〕〔さえ〕〈ない〉。

（B）動詞を棄却して副詞を選択（動詞ＶＳ動詞ルール）
【概略】前の動詞（連用中止）に何も係っておらず、前の多品詞語に副詞があり、前と後の間に読点がなく、次の多品詞に接続助詞や助動詞があり、共起関係がある場合、前の動詞を副詞に交換。
【適用例文】
〔あまり〕〔大きな声で笑う〕〈ので〉、はずかしかった。
〔たとえ〕〔彼がい〕〈ても〉無意味だ。
〔あまり〕〔よく〕〈ない〉。

表２にルールの主な交換パターン(交換前の品詞と交換後の品詞の組み合わせ)と証券17818文でのそのパターンの出現回数を示す。

表２：主な交換パターンと証券文(17818文)での出現回数

交換パターン	例	出現回数(回)
動詞⇔格(副)助詞	さえる⇔さえ	５３５４
動詞⇔名詞	あたる⇔あたり	３８４８
動詞⇔助動詞	続ける⇔続ける	１７０３
動詞⇔副詞	あまる⇔あまり	８５４
名詞⇔単位名詞	円⇔円	５２６０
名詞⇔副詞	ただ⇔ただ	１６５３

表３にルールを導入した場合としない場合の解析の流れの違いを示す。この例においてはルールを導入した場合は４回、導入しない場合は８回で終了する。この例文は実際の文に比べると短く多義・多品詞数も少ないので効果は小さいが、長文の場合に文頭近くでルールが適用されると効果は非常に大きくなる。尚、ルールを導入しない場合、実際には「新聞」と「読む」で依存関係が成立(対象格)するので、(3)で「係る」と判定し、この後「ない」もまとまって誤った解を出力してしまうが、ルールを導入することでこれを回避できる。

表３：「新聞さえ読まない」の解析の流れ

形態素の並び

新聞	さえる	読む	ない
	さえ		

【多義多品詞選択ルールを導入しない場合】
(1)[新聞]⇔[さえる]　:係らない
(2)[さえる]⇔[読む]　:係る
(3)[新聞]⇔[さえる＋読む]　:係らない
（実際には対象格で係るがこの例では係けないでおく）
(4)[さえる＋読む]⇔[ない]　:係る
(5)[新聞]⇔[さえる＋読む＋ない]:係らない（文末で交換）
(6)[新聞]⇔[さえ]　:係る
(7)[新聞＋さえ]⇔[読む]　:係る
(8)[新聞＋さえ＋読む]⇔[ない]　:係る（終了：正解）

【多義多品詞選択ルールを導入した場合】
(1)[新聞]⇔[さえる]　:係らない（交換）
(2)[新聞]⇔[さえ]　:係る
(3)[新聞＋さえ]⇔[読む]　:係る
(4)[新聞＋さえ＋読む]⇔[ない]　:係る（終了：正解）

４．評価実験

以上のようなルールをインプリメントしたシステム(Ｃ言語で記述、構文解析全体で約35Kステップ)で評価試験をSun_SS2(SPECint=21.8,mem=32M)で行ったが、図２に、無作為抽出21423文を実行し、多義多品詞選択ルールを導入した場合としない場合の平均実行時間(実測値)を点ａ、ｂで示す。本システムは最初の解を正解とするが、参考までに全解探索した場合の実行時間も点Ａ、Ｂで示しておく。但し、Ｂについては実行時間が膨大になることから、2000文のみを実行した結果を若干補正して表示した。

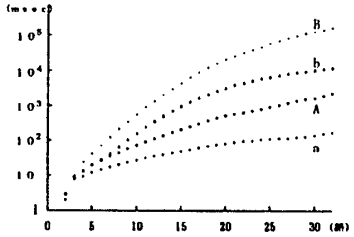

図２：無作為抽出21423文の語数毎の平均実行時間

表４は、上記結果の特定の語数における具体的な平均実行時間と、多義多品詞選択ルールありで実行したときの最初の解を正解とした場合に、ルールが無ければ何個の誤った解候補を出力しうるかを示したものである。

表４：実行時間(msec:括弧外)と解候補数(個:括弧内)

実行種別＼語数	10(語)	20(語)	30(語)
a:ルール有(単解)	28.4(1.0)	84.8(1.0)	143.3(1.0)
b:ルール無(単解)	166.1(3.0)	3328.8(23.3)	10778.5(43.5)
A:ルール有(全解)	75.7(2.9)	513.0(7.5)	1698.5(11.1)
B:ルール無(全解)	566.1(8.1)	20817.4(160.2)	130443.2(463.1)

図２と表４から、実行時間や解の候補数が大幅に低減されていることがわかる。但し、多義多品詞選択ルールがある場合でも、全解探索した場合には、必ずしも正解を１つに絞り込めていない(30語の場合、平均11.1候補残る)。尚、試験に用いた辞書(約４万語)は無作為抽出文に対する多義多品詞語登録が比較的少ないため、探索空間も小さ目であったが、多義多品詞語が多くなれば、ルールの効果はより大きくなると考えられる。多義・多品詞語を比較的多く含んだ長文の具体的実行例を示しておく。

【例文】決算期末を間近に控えた機関投資家の多くは「これが本格的な反騰に結び付くといえるほど外部環境が好転したわけでなく、へたに手を出してここで新たな損は抱えたくない」（大手生保）と慎重な判断を示していた。
⇒語数[53]、形態素数[175]、解析時間[0.148]秒

【例文】関税貿易一般協定・多角的貿易交渉（ガット・ウルグアイ・ラウンド）の農業交渉で、米国と欧州共同体（ＥＣ）が国内農業の保護水準を三〇％に削減する期間を、日本の主張の半分（五年間）に短縮することで合意する見通しが強まったためで、農水省は保護削減強化に対応してこれまでの政策価格（価格支持）方式の一部を農家への所得補償方式に切り替える方向で検討する。
⇒語数[80]、形態素数[378]、解析時間[0.355]秒

５．おわりに

本稿では、多義多品詞選択ルールを利用した依存構造解析について述べた。しかし、ルール内容の詰めはまだ甘く、理論的な根拠にも乏しい。今後は、多義多品詞選択ルールのより一層の拡充による高精度・高速化とともに、多義多品詞選択ルールの理論付けや横型探索方式との融合なども課題として作業を進めていきたいと考えている。

[参考文献]
[1]長尾真:言語の機械処理,三省堂,1984
[2]田中穂積:自然言語解析の基礎,産業図書,1989
[3]田中穂積:自然言語の高速な構文解析法,電子情報通信学会誌,Vol77,No.10,1994
[4]田中,他:先祖表を利用した一般化ＬＲパージングアルゴリズムのファミリー,情報処理学会、自然言語処理90-5,1992