自然言語処理 132－11
(1999. 7. 23)

規則と用例を用いた構文意味融合型日本語構文解析

高橋博之 宮崎正弘

新潟大学大学院自然科学研究科
〒950-2181 新潟市五十嵐二の町 8050
TEL: 025-261-2502
E-mail: { hiro, miyazaki } @nlp.ie.niigata-u.ac.jp

あらまし 日本語の長文の構文解析においては、高精度の従属節間の係り受け解析が必要である。本論文ではすでにある程度の成果を上げている従属節の階層分類による規則ベースの解析手法を単純化し、それにコーパスからの統計情報を組み合わせて解析する手法を提案する。コーパスから統計情報の抽出においてはその分類基準を表記とすることで、分類の恣意性を排除するとともに、意味的な流れの情報を取得し、形式的分類による規則ベースの手法とは異なるアプローチになっている。このように規則解析を構文的制約として、統計解析を意味的制約としてそれぞれ用い両者の統合によって解析精度を向上させることを可能とした。

キーワード 係り受け解析, 日本語従属節, 統計的解析, 規則解析

Integrated Japanese Dependency Analysis using Rules and Corpus

Hiroyuki Takahasi Masahiro Miyazaki

The graduate school of Science and Technology, Niigata University
8050, Ikarashi 2-no-cho, Niigata 950-2181, Japan
TEL: +81-25-261-2502
E-mail: { hiro, miyazaki } @nlp.ie.niigata-u.ac.jp

Abstract It is necessary to improve dependency analysis of Japanese subordinate clauses for precise analysis of long sentences. This paper proposes a integrated approach using rule-based method and corpus-based statistical method. As for extraction of statistical information from corpus, we don't use any classification of clauses preventing risk of biased classification ,but uses its appearance directly. We integrate rule-based method and corpus-based statistical method, selecting suitable method between two methods according as raliability of each result. Using this method, the experment shows that accuracy of the dependency relation between clauses is improved.

key words dependency analysis, Japanese subordinate clause, statistical analysis, rule-based analysis

1 はじめに

日本語の長文の構文解析においては、従属節間の係り受けの誤りが大きなボトルネックとなっている。従属節の係り先の解析誤りはその従属節周辺の格要素などの係り先の誤りを誘発しやすく、それによって文の構造を大きく損ねてしまう。その一方で従属節の係り先を正しく決定することができれば、その周辺要素の係り先も比較的容易に決定することができる。

従来、日本語の従属節の係り受けに関する研究としては南 [1, 2] による従属節の階層分類がよく知られており、白井ら [3] はこの階層分類を計算機による解析での効率性という観点で再整理したうえで、これを詳細化した手法で高精度の従属節係り受け解析を実現している。しかしこの精度は人手による例文の詳細な分析とそれに基づく節の分類に依存しており、このような手法は網羅性に欠けるおそれがあるため、実際に自動処理で多様な文を処理したときに高精度な解析ができるかは疑問である。

これに対して、最近ではコーパスからの統計情報を利用した従属節の解析手法がいくつか試みられている [4][5]。これらの手法は人手の規則では網羅性や保守性に問題があるとして、これを統計情報で置き換えることを目的としている。しかしこれらの統計的手法でも、コーパスの各従属節に与える素性の集合は人手で設定したものであり、しかも両者ともこの素性の種類が約300種類と非常に多くなってしまっており、これも網羅性、保守性の観点から望ましいとは言えない。

規則は詳細になり過ぎると保守しにくくなるが、比較的単純な規則ならば保守しやすい。しかし単純な規則では各種の例外に対処できないためどうしても精度に限界がある。一方、各種の例外的な表現に網羅的に対処するという点では統計的手法の方が有利である。

そこで本論文では従来の規則を統計情報で置き換えるのではなく、規則は単純化した形で基盤として残し、それと統計的手法を組み合わせることで、保守性と網羅性に優れた解析を実現する手法を提案する。

規則としては白井ら [3] の手法を踏襲しつつ、節の分類が基本13種、再分類4種であるのを、11分類に止め、恣意的な要素の強い従属節の分類手法を単純な規則に置き換えて単純化し保守を容易にした。

統計情報としては、素性を与えて頻度集計するのではなく、表記レベルでの集計とすることで、分類の恣意性を排除し、大量のコーパスを自動解析してその結果から情報を収集することで網羅性を確保している。

規則と統計的手法との融合方法としては、それぞれの手法での正解の分布を調べることで、信頼性パラメータを設定し、その比較により信頼できる方の解析結果を採用するという手法をとった。

最後に評価を行い、二つの手法を組み合わせることで、それぞれの手法を単独で使用するよりも高い精度で解析できることが示された。

2 規則による解析

従属節の係りに関しては白井ら [3] が従属節の分類を用いた解析手法について詳しく論じている。本節ではまず、この手法について述べ、次にその規則の単純化について述べる。

2.1 従属節の分類

白井らは日本語の階層的な認識構造に着目し、南 [1, 2] が行なった従属節の意味的な分類を計算機で処理しやすいように以下の3つのレベルに再分類している。

A 類　「同時」の表現
B 類　「原因」「中止」の表現
C 類　「独立」の表現

ここで、この3レベルの節の間にはA類＜B類＜C類、という関係が成り立つ。ここでX＜Yは係り受けにおけるXの優先度がYより小さいということを意味する。つまり、X＜YであるXはYを越えて係ることができず、またYはXに係ることはなく、常にXを越えて係ることになる。

係りの越える越えないには読点の有無が強く影響する。そこでA, B, Cそれぞれを読点の有無で2つに分け、優先度の強さを

A類＜「A類+読点」＜B類＜「B類+読点」
＜C類＜「C類+読点」

としている。さらに、B類については表現の中止性の強弱で2つに分けている。

ここまでは主に付属語表現に関わる部分であるが、これとは別に述語の動作性・状態性に着目した4種の細分類を導入している。

また、引用節が連用節を受ける場合は「C類+読点」相当とし、連体節は形式名詞に係るものは「B類+読点」相当、それ以外はB類相当としている。

2.2 規則の単純化

[3] では節の分類の例を示しているが、その分類の明確な基準は示されていない。また「ことを含め」のように語尾表現を長単位で抽出しているが、一つの節としてまとめる範囲を自動で抽出するのは困難である。そこで、我々は表1に示すような形式的な分類を導入した。ここでの接続助詞の分類は宮崎ら [6] による品詞分類での「同時」「条件」「展開」の分類に相当する。

表 1: 従属節の分類基準

A類	接続助詞 「し」「つつ」「ながら」
B類	接続助詞「ば」「とも」「とて」「ても」「でも」 連用中止形 用言の仮定形 体言止め（「～した結果、」など）
C類	接続助詞 「と」「なり」「が」「けれども」「けれど」「けど」「けども」「けど」「に」「から」「のに」「ものの」

B類に関する中止性の強弱については基準が明確に設定できないため使用しない。また、係り対の述語の種類に関する制約は、後述する統計的手法で補えると思われるので、述語の種類による4つの細分類も単純化のため行わないこととする

引用節と連体節の扱いは白井らの手法と同じである。

なお、我々は従属節以外の要素もこの階層分類に取り込むことで、単純な規則で解析を行うパーザを開発している（規則の詳細は付録参照）。このパーザは後述の統計情報を得るための自動解析に使用し、また、統計情報との融合のための基礎として使用している。

3 統計的情報による解析

コーパスからの統計情報を利用した従属節の係り受け解析手法としては西岡山ら [4] と河原ら [5] によるものがある。いずれの手法でもコーパス中の従属節に素性を与えて、その素性ごとの頻度集計という手法をとっているが、これらの素性は人手で設定したものであり、またその種類が約300種類と非常に多くなっているため、網羅性、保守性の観点から望ましいとは言えない。

従属節の係り対をその表記で見ると、「するには... 必要だ」というような特定の意味の流れに対応する定型表現がしばしば見られる。したがって、このような表記パターンの出現頻度を集計することで、意味の流れのパターンを抽出できるものと考えられる。そこで、我々は恣意性の入らない情報抽出法として節の表記レベルで集計する方法を試みた [7]。

この集計に使用するコーパスは節の係り先が解析してある解析済コーパスである必要がある。集計が字面で行われることから、十分な頻度を得るためには大量の解析済コーパスが必要であるが、現在利用できる解析済コーパスは限られている。そこで、前述の規則解析の手法でコーパスを自動解析し、その結果からの集計を行った。

集計対象は、「するには」-「必要だ」というように、係り元の節末表現と係り先の用言のペアで、それぞれの表現ごとにその出現頻度を集計する。ただし係り元はどの部分までが定型表現なのかは特定できないので、重複して集計する。例えば「実現するには」-「必要だ」という係り対があったら、「実現するには」-「必要だ」、「には」-「必要だ」、「は」-「必要だ」のそれぞれのペアに頻度1を加算する[1]。

集計された頻度は係り元、係り先の出現確率で正規化する[2]。これを相関指数と呼ぶ。この指数が大きいほど係り元-係り先の相関が強い。

この統計情報を使った解析では、係り元と各係り先候補との間の相関指数を求め、一番高いものを係り先に採用する。

集計には日経新聞の94年の全文記事[3]を用いた。集計の規模を図1に示す[4]。

4 規則解析と統計解析の融合

規則と統計情報を併用する一つの方法はそれぞれの手法で求められた評価点を足し合わせて、あるいは掛

[1] 係り元は単語境界以外では切らない

[2] 相関指数 $= \frac{\text{係り対の出現確率}}{\text{係り元表現の出現確率} \times \text{係り先表現の出現確率}}$

[3] 「日経全文記事データベース日本経済新聞 CD－ROM版94年版」を使用

[4] ここに示した数値は [7] の集計での不具合を直して再集計した時のものである

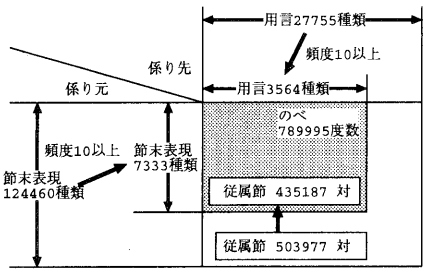

図 1: 集計規模

け合わせて求めた総合評価点を比較して係り先を決定する方法である。しかし、もともと別々の基準で得られた評点を組み合わせるためにはやや恣意的な数式の調整が必要になり、根拠が明確でなくなるおそれがある。

そこで、それぞれの手法についてその解が正解である確率を示す信頼性パラメータを算出し、その比較によってどちらかの手法を選ぶことにする。例えば規則手法での解析結果の信頼性が 80% で統計的手法での解析結果の信頼性が 90% ならば、統計的手法での結果を採用する。当然、このような信頼性の比較は双方の結果が異なった場合にのみ行う。

4.1 規則解析の信頼性

優先度に基づく規則解析手法では係り元と係り先候補との優先度が近接しているほど、係る／越えるの判定の誤りの可能性は増えると予想される。実際に、後述する評価結果から優先度の差と正解率との間の相関を見てみると図 2 のように優先度の差と正解率はほぼ比例しており、優先度の差が大きいほど正解率は向上している。

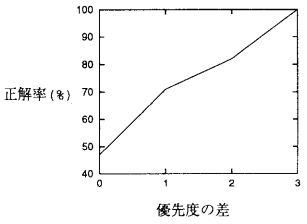

図 2: 規則解析の信頼性

次に、一般的傾向とは別に節の種別による精度の違いについて調査した。図 2 でわかるように、優先度の差が 0、つまり階層分類での同じレベルの場合に係るか越えるかの判定が最も難しくなるが、特に形式名詞に係る連体節にその前の連用節が係るかかどうかの判定で誤ることが多かった。[3] では形式名詞に係る連体節は対象をとらえ直すためにそれを越えることはないとしているが、「～会談し、～することで～合意した。」のように、越えた先に意味的につながりやすい用言があるような場合には形式名詞に係る連体節を越えて係る場合もある。このような場合にはコーパスからの意味の流れの情報が効果的に働くと思われるので、統計解析を利用したほうが高い精度が期待できる。

4.2 統計解析の信頼性

統計解析では統計情報を基に、係り元の節と各係り先候補の節との相関指数をそれぞれ求め、その比較で係り先を決定する。相関指数は意味的相関の強さを示すものであり、各候補間でその差が大きくなった場合の方が正解率が高いと予想される。実際に、後述する評価結果から相関指数の差（倍率）と正解率との間の相関を見てみると図 3 のように相関指数の差と正解率はほぼ比例しており[5]、相関指数の差が大きいほど正解率は向上している。

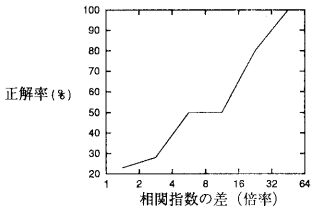

図 3: 統計解析の信頼性

5 評価

評価は日経新聞の記事[6]から抽出した 493 の従属節について行った。なお、処理を単純にするために、ここ

[5]相関指数の倍率で問題を 6 クラスに分け、それぞれの中での正解率を求めた

[6]集計に使用した文とは別のもの

では係り先候補が2つである従属節に限定し、係り先がどちらとも取れるようなものはできる限り排除した。

まず、規則解析手法と、統計的解析手法をそれぞれ単独で使用した場合の正解率を求めた。これを表2に示す。

表 2: 各手法単独での評価結果

解析手法	正解数	正解率
規則解析	385	78%
統計解析	310	63%

次に双方の解析手法の選択基準を定める。まず、図2,3に示した規則解析と統計解析の信頼性の傾向を線形回帰計算（残差最小法）で近似して以下の式を得た。

$$\text{規則解析の信頼性 (\%)} = 18 \times \text{優先度の差} + 46$$

$$\text{統計解析の信頼性 (\%)} = 18 \times log_2(\text{相関指数の差})$$

この式に基づいて双方の手法を選択的に使用する。ただし、前述のように形式名詞に係る連体節のケースでは常に統計的手法を用いる。この場合の結果を表3に示す。

表 3: 融合手法の評価結果

		該当する節	正解 (正解率)
解が一致		302	252(83%)
解が不一致	規則解析を選択	163	125(77%)
	統計解析を選択	28	20(71%)
合計		493	397(81%)

最終的に、規則のみで解析した場合（表2）よりも約3% 精度が向上している。あまり精度が向上していないのは、統計的手法の精度がまだあまり高くないために、統計的手法を利用できるケースが少ないためであると思われる。これは統計が自動処理によっているために、集計ミスが含まれていること、集計の対象が係り元の節末と係り先用言に限定されていることによる情報不足などが原因であると推定される。

統計解析は全体では正解率は 63 % でしかないが、融合手法で採用されたケースでは 71 % と、10 % 近くの向上になっており、信頼性の高い解を採用するという方式がうまく働いていることがわかる。

また、両方の解が一致した場合にはより高い精度が得られており、これは今後、コーパスから自動解析で情報収集をする際に役に立つと思われる。

6 おわりに

規則を主に構文的制約に、コーパスからの統計情報を意味的制約に使い、それらを選択的に使用することで、双方の制約を対立させること無く融合させ、それぞれを単独で使用するより高い正解率を実現した。

統計的手法の精度向上が今後の課題である。

参考文献

[1] 南不二男. 述語文の構造, 日本の言語学. 大修館書店, 1964.

[2] 南不二男. 複文, 講座現代語 6. 明治書院, 1964.

[3] 白井諭, 池原悟, 横尾昭男, 木村淳子. 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度. 情報処理学会論文誌, Vol. 36, No. 10, pp. 2353–2361, 1995.

[4] 西岡山滋之, 宇津呂武仁, 松本裕治. コーパスからの日本語従属節係り受け選好情報の抽出. 情報処理学会自然言語処理研究会 126-5, 1998.

[5] 河原大輔, 黒橋禎夫. 従属節に関する統計的情報と一般的統語規則を結合した日本語構文解析システム. 自然言語処理学会第5回年次大会発表論文集, pp. 536–539, 1999.

[6] 宮崎正弘, 白井諭, 池原悟. 言語過程説に基づく日本語品詞の体系化とその効用. 言語処理学会誌, Vol. 2, No. 3, pp. 3–25, 1995.

[7] 高橋博之, 宮崎正弘. 大規模コーパスを用いた日本語従属節パターン抽出. 自然言語処理学会第5回年次大会発表論文集, pp. 546–549, 1999.

付録—使用した解析規則—

表 4: 接続優先度

優先度	節や句の種類
6	主節 C 類の節＋読点 「引用」の節＋読点 接続詞＋読点
5	C 類の節 「引用」の節 接続詞 長距離性格後置詞句（「... は」）＋読点
4	B 類の節＋読点 形式名詞に係る連体節（受け側） 長距離性格後置詞句（「... は」） 格後置詞句＋読点 名詞並列句＋読点 副詞句＋読点
3	B 類の節 普通名詞に係る連体節（受け側）
2	C 類の節＋読点
1	C 類の節 格後置詞句 名詞並列句 副詞句 連体節（係り側） 形式名詞＋「の」 形式名詞に係る「名詞＋の」 名詞＋読点（並列の用法）
0	名詞＋「の」（上記の優先度 1 となるものを除く） 連体詞

表 4 は従属節の 6 レベルの優先度分類を従属節以外の句にも拡張したものである。係りに関する規則は従属節の場合と同様で、各句や節は自分と優先度が同じかより高い句や節には係るが、より小さい優先度の句や節には係らない。ただし、連体節は自分が係り元となる場合と係り先になる場合とで優先度が異なる。

図 4 に接続優先度の使用例を示す。「おける」が自分より優先度の高い「画期的な」を越えて係っているが、これは連体節の係り先が名詞（名詞句）に限られるためである。

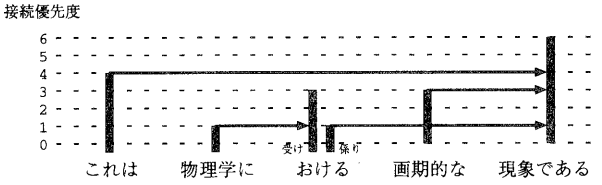

図 4: 接続優先度の例 (1)