大規模コーパスを用いた日本語従属節パターン抽出

高橋博之　　宮崎正弘

新潟大学大学院自然科学研究科

1 はじめに

従属節の係りの解析手法としては、例えば「○○して」は「○○しつつ」を越えるというように従属節の付属語部（節末表現）の組に対して「越える／越えない」の規則を用いる手法が提案されている [1]。

しかし、このような規則は主に付属語表現についてのもので、用言の情報をほとんど使わず、文の意味の流れを反映していない。例えば「○○するには」という「目的」の用法の従属節は意味的に「必要だ」や「不可欠だ」などの用言に係りやすいが、このような情報は付属語表現に着目した規則では記述できない。

意味の流れという観点からみると、係り元の節末表現と係り先の用言との間に関連性がある場合が多い。そこで、この節末表現と用言との間の関連性を規則化し、解析などに利用することを考えた。

このような意味情報は人手で網羅的に記述するのは困難である。そこで、コーパスを用いて自動抽出する。コーパスとしては従属節の係り対を取り出せる解析済みコーパスが必要であるが、現状では利用可能な解析済みコーパスは多くない。そこで、すでに開発済みの日本語構文解析システム Jip を用いて大量のコーパスを自動で解析させ、その結果から従属節の係りパターンを抽出した。

抽出には新聞記事から 70 万文を使用した。抽出した各係り対パターンにはその頻度を元に、意味的相関の強さを示す指数を付与した。

最後に抽出されたデータの意味的検証のために、簡単な解析実験を行なった。その結果、有効に適用できる文は多くはないものの約 9 割の正解率が得られ、データの精度の高さが示された。

Extraction of Japanese Phrase Pattern using Large-scale Corpus.

Hiroyuki Takahashi (hiro@nlp.ie.niigata-u.ac.jp),
Masahiro Miyazaki (miyazaki@ie.niigata-u.ac.jp)
Niigata University

2 従来の手法の解析精度

従属節の係り先の解析手法としては、[1] が従属節の階層分類を利用した方法を提案している。[1] ではこの手法による従属節単位の正解率が 99.3 % とされているが、この結果は手動処理でのもので、[2] による自動処理での追試では 81 %という数値が出ている。ただし、後者は係り先が自明な主節の直前の従属節を母集合に含めていない。

我々が開発した日本語構文解析システム Jip は従属節の係り先解析手法として、[1] とほぼ同様の手法を用いている。Jip を用いて行なった解析実験では 260 文を使用して従属節の係りの正解率は 87% で、係り先の自明なものを母集合から除くと 77.5 % [1]であった。これは [2] での 81% という数値に近く、[1] の手法での解析精度の一応の基準となる。

今回の従属節パターン抽出では Jip で解析した従属節の係り対を元データとして使用した。すなわち元データの約 1 割が誤りである。

3 抽出する情報

従属節の係り対では「○○するには...必要だ/で○○」というように係り元の節末表現と係り先の節頭表現が固定されたパターンが多く見られる。この固定された部分に典型的な意味の流れがあると考え、こういったパターンをコーパスから抽出する。

[1]従属節 225 のうち 196 が正しく係り先を求められたが、そのうち 96 は係り先が自明であった

3.1 係り元の固定部分

係り元の文節での固定部分はその末尾の部分であるが、これは必ずしも付属語の部分だけではない。例えば「比べて...大きい」では動詞「比べ」を含んだ「比べて」全体を固定要素と見なすべきである。このように係り元の固定部分をその品詞で特定することはできない。

そこで、後述するように全ての可能性について重複して抽出・集計し、実際に使用する際に意味的な関連性の最も高いものを使用することとした。

3.2 係り先の固定部分

係り先の文節の固定要素はたいていが、先頭の用言[2]と見てよい。そこで用言のみを抽出し、用言の後の接尾語・助動詞等は無視する[3]。用言自体の活用形も無視する。

4 抽出・集計方法

コーパスから自動解析で集めた従属節の係り対から係り元の節末表現と係り先の用言の対を抽出し、その出現頻度を集計する。

前述のように係り先の固定部はその用言であり一意に決まるが、係り元の節末の固定部の範囲は特定できない。そこで、全ての節末表現、つまり末端の単語を含む全ての部分単語列について重複して頻度を集計する。

例えば、「実現するには... 必要である」という係り対があったら、図1のように「実現するには」-「必要だ」、「には」-「必要だ」、「は」-「必要だ」の3つの対に頻度1を加える。なお、読点は意味的流れとは関係がないので節末表現に含めない。

4.1 正規化

頻度集計の後にはその正規化処理が必要である。これは意味的相関が低いパターンでも単にそれぞれの表現の出現頻度が高ければその組合せの出現頻度も高くなってしまうからである。

[2]ここでは動詞、形容詞、形容動詞だけでなく、名詞+「だ」と体言止めも用言に含める

[3]「には ...欠かせない」のように後接の助動詞等を含めて固定される場合もあるが、このようなケースはあまり多くないので今回は無視した。

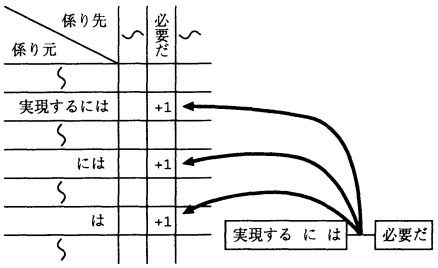

図 1: 重複集計の例

例えば図2に示すように、「には」-「ある」と「には」-「必要だ」の出現頻度はほぼ同じである。しかし、それぞれの条件なし出現確率[4]をみると、用言「ある」は「必要だ」よりも約 18 倍出現しやすいので、この頻度をそのまま比較しても意味はない。そこで、それぞれの出現確率で割る必要がある。同じことは係り元の節末表現の出現確率についても言える。

係り先 / 出現確率 / 係り元	ある	必要だ
	0.039	0.0022
には	178	195

図 2: 正規化の必要性

また、集計に使用した従属節対の数で割ることで、集計に使用した従属節対の数と関係のない指数になる。

以上を整理すると、節末表現 i と用言 j の対の正規化された頻度 $C(i,j)$ は

$$C(i,j) = \frac{F(i,j)}{N * P_1(i) * P_2(j)}$$

となる。ここで、$F(i,j)$ は節末表現 i と用言 j の対の出現頻度、N は集計に使用した従属節対の数、$P_1(i)$ は節末表現 i の条件なし出現確率、$P_2(j)$ は用言 j の条件なし出現確率、である

この正規化された頻度 $C(i,j)$ は節末表現と用言の相関を示すものであると考えられる。すなわち、$C(i,j) = 1$ ならば特に相関はなく、1 より大きければ正の相関が 1 より少なければ負の相関があると考えられる。この $C(i,j)$ を相関指数と呼ぶ。

[4]係り元に関係のない出現確率

5 抽出実験

日経新聞の 94 年の全文記事[5]から、長い文、引用符を含む文など、解析が失敗しやすそうな文を除いた約 70 万文を使用し、これを自動解析させて従属節対[6]約 49 万を得た。

5.1 抽出する範囲

集計時には節末表現と用言を縦横にとった 2 次元配列が必要となる。この配列の大きさを見積もるために予備調査を行った。その結果用言は全部で約 3 万 2 千種類[7]、節末表現は重複集計で約 9 万 6 千種類であった。

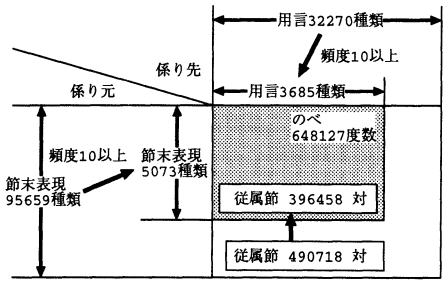

図 3: 集計規模

これだと配列が大き過ぎるので、頻度制約を設けることにした。用言は頻度 10 以上に限定し 3685 種類とした。これで従属節対全体の約 9 割をカバーする。節末表現も頻度 10 以上に限定し 5073 種類とした。従属節全体の約 9 割がこの内のいずれかの節末表現を含む。この制約により、実際に利用できた従属節対は約 8 割の約 40 万対である。ここから前述の重複集計により約 65 万度数が加算された。図 3 に集計規模を示す。

5.2 抽出されたパターンの例と問題点

抽出されたパターンのうち相関度の高いものの例を表 1 に示す。

ここで「には」-「必要だ」と「には」-「一番だ」は頻度では前者が大きく上回っているが、正規化によって相関指数では逆転している。これは「必要だ」の用例が 1088 あるのに対し、「一番だ」の用法が 11 例しかないため、頻度の重みが大きく異なるためである。このように正規化によって用言の出現頻度の差が補正されている。

表 1: 抽出されたパターンの例

種別	頻度	相関指数	係り元節末	係り先用言
助詞	6	27.87	より	安い
	195	39.90	に は	必要だ
	3	60.71	に は	一番だ
	35	285.94	か は	微妙だ
形容詞	2	3015.73	温かく	迎え入れる
	4	3285.22	手厚く	保護する
	3	4007.06	慌ただしく	出入りする
動作の連携	4	3250.64	結婚し て	産む
	3	3327.70	評価され て	受賞する
	2	4595.40	包ん で	揚げる
	5	5676.67	隠し	脱税する
反復表現	4	5146.85	勝てば	勝つ

助詞部分の節末表現では十分な頻度が得られていることが多いが、係り元用言まで含めた節末表現では頻度は 2,3 程度が多く十分とは言えない。頻度が少ないと頻度自体が誤差を含むし、解析誤りの影響をうけやすい。また、用言を含めた節末表現はそれ自体の出現確率が小さいため、頻度の多少の違いが相関指数に極端に反映されてしまう。例えば「結婚し」-「産む」の 3250.64 に対し、「結婚し」-「生まれる」は 228.28 と意味的つながりはさほど変らないものでも相関指数が極端に違う例が見られた。

こういった問題の解決法の一つは入力文例を増やす事である。このデータは自動解析で得られたものなので、電子化された文書と時間さえあればいくらでもデータを増やすことができる[8]。

入力文例を増やす他に、似た用言をまとめることで、頻度を増やす方法も考えられる。ただし、その場合はどこまでを類似とするかの基準が問題となる。

また、今回は新聞記事を使用したため、「運ばれたが」-「死亡する」のような新聞記事ならではの表現に頻度の偏りがやや見られた。これは今後新聞以外の文献を使用することで対処できると思われる。また、自動処理の利点を生かして、分野別のデータを用意して、解析対象によって使い分けることも可能である。

[5]「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

[6]従属節と主節の対を含む

[7]「１０００倍だ」のような数詞＋「だ」を表記で分けてしまったためかなり多くなった

[8]コーパスの解析は一文 5 秒ほどで 24 時間で 1 万 7 千文を処理できる。今回使用した 70 万文の解析は WS(SparcStarion 5) を 5 台用いて 1 週間ほどで完了した。

6 解析実験

抽出実験で取り出されたデータの精度を検証するため、簡単な解析実験を行なった。

ここでは処理の単純化のため「節 A... 節 B... 節 C (文末)」という節3つの形式の文を集め、節 A の係り先（節 B か節 C か）をどのくらいの精度で求められるかを調べた。

6.1 解析対象

解析実験の対象としては日経新聞 94 年の全文データのうち、前述の抽出実験に使用しなかったものを用いた。ここから「節 A... 節 B... 節 C (文末)」という形式の文を 500 文集め、手作業で節 A の係り先を決定し、正解データとした。ただし、「川には三十石船が浮かび、堤を人々が往来し、柳の木の下に茶店があった。」のように節 A の係り先が節 B か節 C かはっきりしないものは除いた。

6.2 係り先の決定法

以上のようにして収集した「節 A ...節 B ...節 C (文末)」という形式の文について、節 A の節末表現と節 B の用言との相関指数、節 A の節末表現と節 C の用言との相関指数をそれぞれ求め、相関指数に有意な差が出た場合に大きい方に係るとした。有意な差とは R 倍以上の差とし、この R を変えて実験した。

前述のように節末表現は重複集計であり、複数の節末表現-用言パターンと一致するケースがありうる。その場合は一番相関指数の高いものを採用することとした。

係り先の決定には相関係数が求まることが前提である。そのためには以下の条件を満たす必要がある。

1. 節 A の節末表現のいずれかが集計対象に入っていること
2. 節 B の用言が集計対象に入っていること
3. 節 C の用言が集計対象に入っていること

解析の手法としては節末表現、用言ともに類似のもので代用するという方法も可能であるが、今回はデータの内容の検証が目的なので完全一致のみとした。

また、前述のように出現頻度の小さい場合は相関指数に誤差が出やすい。そこで、全てのパターンの相関指数を使用した場合と、出現頻度が 2 以上のパターンの相関指数のみ有効とした場合で実験した。

6.3 結果と分析

解析実験の結果を表 2 に示す[9]。頻度制約を付けない場合の正解率が 80% 弱で頭うちなのは、元データに含まれる誤りの影響が出ているものと思われる。頻度制約を加えた場合は、R の値が増加するに従い正解率も向上する。また、実際にマッチした文例を見てみると、$R = 2$ 程度では「○○て」「○○が」「○○ば」など、特に用言を選ばない短い表現にマッチすることが多く、パターンの効果があまり現れていない。$R = 4$ 以降では「派遣して」-「指導する」のように意味的流れが明白な長めのパターンにマッチすることが多く、ここでは 9 割程度の正解率を達成している。

表 2: 解析実験結果

	倍率 R	1	2	4	8	16
頻度制約なし	正解率	64%	70%	77%	79%	77%
	カバー率	67%	45%	32%	26%	22%
頻度 2 以上	正解率	62%	70%	85%	92%	97%
	カバー率	46%	26%	14%	10%	6%

7 おわりに

大規模コーパスから自動抽出した従属節の係り対から節末表現-用言パターンを抽出し、評価によってその精度を確認した。

このパターンデータは意味的流れの明白でない文には適用できないので、形式的情報との組合せが不可欠であり、そのための手法の検討が今後の課題である。

参考文献

[1] 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.

[2] 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.

[9] ここで正解率は $\frac{\text{係り先が正しく求められたもの}}{\text{パターンが適用できたもの}}$、カバー率は $\frac{\text{パターンが適用できたもの}}{\text{解析文数 (500)}}$ である。