E-32　並列構造を加味した係り受けの複雑さの指標の作成と推敲支援への応用

A Complex Dependency Indicator with a Factor of Coordinate Structures and Its Application to a Writing Tool

横林 博† 菅沼 明‡ 谷口 倫一郎‡
Hiroshi Yokobayashi Akira Suganuma Rin-ichiro Taniguchi

1. はじめに

近年、私たちは計算機で文章を作成することが多くなり、またそれらは校正支援機能や推敲支援機能を持つようになった。これらを使えば、効率よく文書を作成することが可能になる。

本研究では一歩踏み込んだ推敲支援情報として、文の係り受けの複雑さに着目した。係り受けが複雑な文には、読みづらい文が含まれていると考えられる。我々はすでに係り受けの複雑な文を抽出するモデルを作成した[1]。このモデルは係り受けの複雑さを表す指標を算出し、その指標を基に文の複雑さを判定している。ただ係り受け解析過程モデルでは並列構造解析を行わず、文中の並列構造を考慮していない。そこで本論文では並列構造を係り受けの複雑さの指標に加味し、複雑な並列構造を持つ文を抽出することを考える。

2. 並列構造を加味した係り受けの複雑さの指標を算出するモデルの作成

2.1 係り受け解析過程モデル

一般的に人間が文を理解する時には文の解析を行っているが、解析作業を行っている間に、人間は解析結果を一時的に保持しておくための短期的な記憶を使用していると考えられる。京都大学の村田らは、短期記憶に格納する必要があるものは係り先が未決定な文節であると考え、京都大学テキストコーパスを用いて調べたところ、係り先が未決定な文節はそれぞれの文に対して平均して 7±2 であったと報告している[2]。そこで文処理で用いられる短期記憶の容量も 7±2 程度であると仮定できる。

これに基づいて作成した係り受け解析過程モデルは入力処理部、係り受け判断部、短期記憶スタックの三つの部分から成り立っている。入力処理部では入力文を文節に区切る。係り受け判断部では短期記憶スタックから pop された文節と、入力文節との間に係り受け関係が成立するかを判断する。短期記憶スタックは人間の短期記憶に相当するスタックで、文の処理過程で係り先が未決定な文節を保持する。その後スタック中のブロックと入力ブロックの間に係り受け関係が成立すると、その二つのブロックを結合し、それを一つのブロックとして短期記憶に push する。この作業を文末まで繰り返す。こうして使用したブロックの最大段数を記録し、係り受けの複雑さの指標として使用する。

2.2 並列構造を指標に加味する必要性

人間が文を理解するに当たって、困難さを感じる構造の一つに並列構造がある。複雑な並列構造があると、書き手と読み手の間の解釈に食い違いが生じやすく、それを認識する過程で困難さを感じる。つまり、並列構造は文の読みにくさに関して何らかの影響を与えていると考えられる。しかし、上記の係り受け解析過程モデルは並列構造を持つ文に対して並列構造の解析を行わず、並列構造を無視して処理を行う。これでは並列構造が指標に反映されない。そこで、係り受けの複雑さの指標に並列構造を反映させ、複雑だと思われる並列構造を持つ文を抽出することを考え、係り受け解析過程モデルを拡張した。

2.3 並列構造を加味したモデルの作成

並列構造の複雑さの要因についてはいろいろと考えられるが、主に以下の 2 点に関して並列構造を評価し、複雑さの指標に反映させる。

(1) 並列要素の長さや並列要素内の係り受けの複雑さによる読みにくさ
(2) 複数個の並列構造の重なりによる読みにくさ

具体的には、並列構造の並列要素ごとの結合を保留し、並列構造の中で最も後ろに位置する並列要素の終点文節が入力されてから結合を行う。人間が並列構造を持つ文を読むときは、どの文節列とどの文節列が並列であるかを認識する必要があるし、終点文節を読んではじめて並列要素の認識が可能であるからである。

しかし上記の処理では問題点も出てくる。明らかに読みづらさを感じないにも関わらず、複雑さを表す指標が大きくなる構造が存在する。例文を以下に示す。

例文 1　キツネ、ムササビ、フクロウ、トンビ、カエル、ザリガニ、カジカ、バッタ、トンボ……。

例文 1 の並列構造は単純な単語列から成り立っており、並列構造の重なりは存在しないため、この文を読むときには困難さを感じない。人間がこの文を読むときには、各単語を単純に短期記憶にブロックとして積んでいるとは考えにくい。しかし、並列要素が長くて複雑な場合は、それをひとまとめにして一つのブロックとして文を読んでいるとは考えにくい。

そこで、上記の中間程度の記憶容量を使用しているのではないかと考え、対数を使用する。対数を用いるとブロックの段数に応じた圧縮が可能で、前置要素の文節の長さを圧縮値に反映させることができる。このアルゴリズムを図 1 に示す。

3. 実験

作成した並列構造処理モデルを評価するために実験を行った。評価には、京都大学テキストコーパスの中で並列構造を含む文 3283 文を用いた。並列構造を加味した係り受けの複雑さの指標と、文の数の関係を表 1 に示す。

† 九州大学大学院システム情報科学府
‡ 九州大学大学院システム情報科学研究院

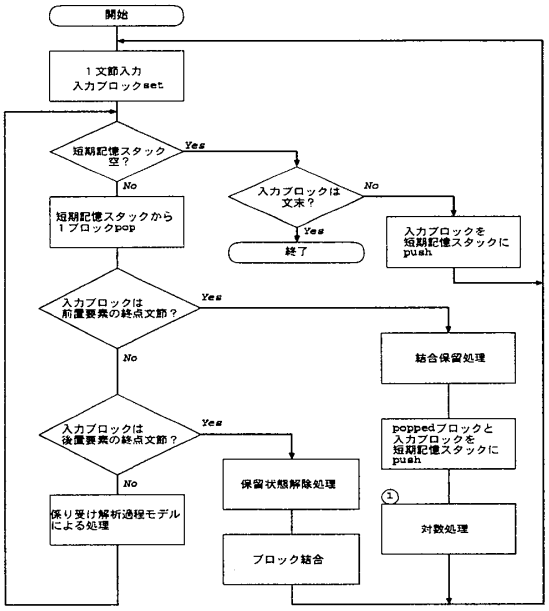

図1　並列構造を考慮したアルゴリズム

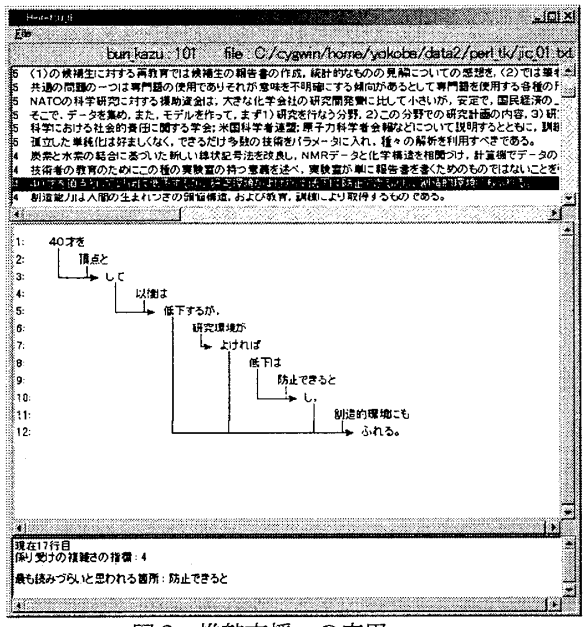

図2　推敲支援への応用

係り受けの複雑さの指標が 7±2 の上限である 9 を超える文は、66 文存在した。その中から例を挙げる(括弧の中の数値は指標を示す)。

例文 2　経団連の豊田章一郎会長は九日、行政改革委員会の飯田庸太郎委員長ら四委員と都内のホテルで懇談、経団連として委員会活動の全面支援を約束するとともに、実効ある規制緩和推進五カ年計画 が三月末までに策定されるよう政府への監視の徹底を求めた。 (15)

例文 2 では文のほとんどが並列構造になっていて、並列構造内の修飾節が続くために指標が大きくなっている。例文 2 のように、並列要素の長さや並列要素内の係り受けの複雑さ、あるいは並列構造の重なりによって読みにくくなっている文を実験において抽出できたことを確認した。

4. 推敲支援への応用

係り受け解析過程モデルでは、7±2 を基準にして文の読みづらさを判断することが可能である。本手法では構文情報のみを用いて文の複雑さを判定するため、モデルが抽出したすべての文が読み手にとって読みづらいということは考えにくいが、提示した文の中から書き手に推敲を促すことは有意義であると考える。

本手法の推敲支援への応用として、係り受けの複雑さの指標の高い文を書き手に提示するシステムのプロトタイプを Windows2000 上に作成した。図 2 にプロトタイプの表示例を示す。システムは 3 つのウィンドウから成り、上から順に、係り受けの複雑さの指標とその文の一覧、構文情報、メッセージを表示する欄である。一番上のウィンドウでは、入力したテキストファイルの各文の係り受けの複雑さの指標を算出した後、指標の高い文から順に提示する。一番上のウィンドウから文を一つ選択すると、下の 2 つの欄に解析結果を瞬時に表示する。真ん中のウィンドウでは文の各文節ごとの係り受け関係をグラフィカルに表している。

また推敲支援に有用な情報として、文の係り受けが複雑になっている原因箇所の提示を行っている。具体的には、係り受けの複雑さの指標を算出する過程で、指標がもっとも大きくなる文節を文中でもっとも読みづらい箇所として書き手に提示する。これにより、書き手は文のどの部分が複雑さの原因になっているかといった具体的な情報を得ることが可能になる。現時点では、この原因箇所を指摘するメッセージは、プロトタイプの一番下のウィンドウに表示している。

表1　係り受けの複雑さの指標と文の数の関係

係り受けの複雑さの指標	文の数	係り受けの複雑さの指標	文の数
15	2	7	270
13	1	6	491
12	11	5	712
11	15	4	676
10	37	3	452
9	72	2	250
8	111	1	182

5. おわりに

本研究では、係り受けの複雑さに並列構造を加味した指標を算出するモデルを作成し、複雑な並列構造を持つ文の抽出を行った。さらに、係り受けの複雑さの指標を推敲支援へ応用する方法に関して述べた。

今後の課題としてはプロトタイプの改良、訂正に役立つ情報のさらなる提示などが挙げられる。

参考文献

[1]小池康弘、菅沼明、谷口倫一郎、“係り受け解析過程モデルを用いた係り受けの複雑さの指標の作成とその推敲支援への応用”、第 14 回情報処理学会九州支部研究会報告、pp.370-377 (2000)

[2]村田真樹、内元清貴、馬青、井佐原均、“日本語文と英語文における統語構造認識とマジカルナンバー7±2”、自然言語処理、vol.6, No.7, pp.61-71 (1999)