自然言語処理 134−10
(1999. 11. 25)

枝分かれするネットワークを用いた格構造解析システムへの頻度情報の導入とその評価

竹内俊行　荒木健治　栃内香次

北海道大学大学院工学研究科

〒060-8628 札幌市北区北13条西8丁目
TEL 011-706-6823
E-mail:{toshi,araki,tochinai}@media.eng.hokudai.ac.jp

我々は構文解析の出来ない会話文などの格構造を解析するシステムとして単語同士の意味的なつながりを用いて日本語単文の格構造を解析する手法を提案した. また, 実際にこれを用いたシステムを構築し, それによって文の解析実験を行ない, その結果を示し, 考察してきた. 本稿では, より高い解析率を得るために, ある単語がどのような意味で用いられやすい, ある名詞がある動詞の何格として用いられやすい, といった情報をネットワーク上のリンクの頻度として表現した. これにより, 使用頻度の高い格構造を表現するリンクを効率よく活性化することができる. また, そのように改良したシステムによって文の格構造解析実験を行ない, 解析率の向上を確認した. そこで, その結果を示し, 本手法の有効性について考察する.

キーワード　格構造解析, 意味表現, 格助詞

Introduction of Frequency Information to Case Structure Analysis System Using Network with Branched Links and Performance Evaluation of the System

Toshiyuki Takeuchi　Kenji Araki　Koji Tochinai

Graduate School of Engineering, Hokkaido University

N13-W8, Kita-ku, Sapporo 060-8628,Japan
TEL (+81-11)706-6823
E-mail:{toshi,araki,tochinai}@media.eng.hokudai.ac.jp

We have proposed the method which can analyzes the case structures for Japanese by the relation between meanings of words. The system using our proposed method can analyze non-grammatical sentences like concersation sentences. We have developed a system based on our method and analyzed sentences with the system. And we have considered their results. In this paper, we introduse frequency into our proposed method in order to obtain a higher correct answer rate of an analysis. The network is activated more efficiently according to the frequency. We developed a system introdused frequency and analyzed sentences with the system. And we show their results, and consider them.

keywords　case structure analysis, meaning expression, case particle

1 はじめに

日本語の意味解析において格構造解析は不可欠であり, また, 格構造を表現するために格助詞は重要な役割を担っている.

従来の意味解析手法 [2][3] では, 文法を基にして文の解析を行ない, 形態素解析, 構文解析の結果を用いてその意味解析を行なうために, 構文解析まで成功しなければ意味解析を行なうことができないなどという問題が存在した. 特に会話文などには構文解析の難しい文が多数含まれており, 構文解析結果を用いずに意味解析を行なえば, 解析の難しい会話文の意味解析を行なうことができる.

そこで我々は, 構文解析を行なわずに単語分けされた文から直接格構造解析を行なう手法として, ネットワークを用いた手法を提案してきた [1]. 本手法は, 助詞欠落文や語順倒置文など非文法的な文の解析に有効であると考えられる.

この既提案システムに, 頻度情報を導入し, ある語がどのような意味で用いられやすいか, ある名詞がある動詞の何格として用いられやすいか, といった情報を表現した. これにより, 使用頻度の高い格構造を表現するリンクを効率よく活性化することができる. 本稿では, そのシステム全体について述べ, このシステムを用いて実際に文の格構造解析実験を行ない, その結果を示し, 本手法の有効性について考察する.

ネットワークを用いた自然言語処理手法として, McClelland [4] と Rumelhart [5] によるものや Walts と Pollack [6] によるものなどがある. [4][5] では, 特徴, 文字, 語をネットワークで結合し, 与えられた文字列を単語として認識する. これを [6] では, 入力, 語彙, 統語, 文脈をネットワークで結合することにより文の意味理解に拡張している. 本手法ではさらにこの考え方を日本語に用いるために, ネットワークの構造を拡張する. 日本語では名詞, 格助詞, 動詞の 3 つの単語によって格関係を表現するために, それぞれの語の概念をノードとするネットワークでは, 3 つのノードを直接結合するリンクを導入することにより, その格関係を明確に表現することができる.

本手法では, 単語同士の意味的なつながりを利用し, 文を解析する. 単語同士の意味的なつながりはネットワークを用いて表現し, このネットワークは, 単語とその表現する概念とをノードとする. また, この手法では格構造解析を行なう際に格助詞を重要視する. 日本語では「名詞-格助詞-動詞」という形で表現される格構造をネットワークを用いて表現するために, ネットワーク上のリンクを枝分かれさせる. 本システムに日本語単文を入力すると, システム内のネットワーク上のノード同士が活性値の相互作用を起こし, 格構造を表現するための枝分かれしたリンクが強く活性化することにより文の格構造解析を行なう. 1 文解析毎にその正誤をユーザが入力し, それによってリンクの使用頻度を更新し, 以後の解析に用いる.

2 システムの概要

システムはネットワークを用いて文の格構造解析を行なう. その処理の流れは図 1のようになる. システムは文が与えられると, その文中の単語に対応するネットワーク上のノードを活性化し, 活性化したノードがほかのノードと相互作用を起こす. 相互作用が収束したとき, そのネットワークの活性化の状態は与えられた文の格構造を表現しており, システムがそれを認識し, 与えられた文中のそれぞれの名詞の動詞に対する格を出力する. また, 1 文の解析毎に解析結果の正誤をユーザに問い, 正しい格構造解析結果に対応するリンクの使用頻度を更新し, それを以後の解析に利用する.

3 格構造解析の手法

3.1 ネットワークと解析の概要

本システムでは, システム内にネットワークを構築し, そのネットワークは単語層と概念層の 2 層に分けられている. その単語層ノードはある単語の表記に対応し, 概念層ノードはある単語の概念に対応する. また, 意味が近いノード同士などの関係の深いノード間にはリンクが張られる.

ノードやリンクは活性値を持ち, あるリンクとその両端のノードの活性値が高いときに, 与えられた文中でのその両端のノードの関係が強いことを表現する. リンク, ノードがともに活性値を持つことにより同様に扱うことができ, 以後, これらをまとめてパーツと呼ぶものとする.

システムに文が与えられると, 与えられた文中の単語に対応する単語ノードが活性化する. 活性化し, 活性値を持ったパーツは隣接するパーツを活性化さ

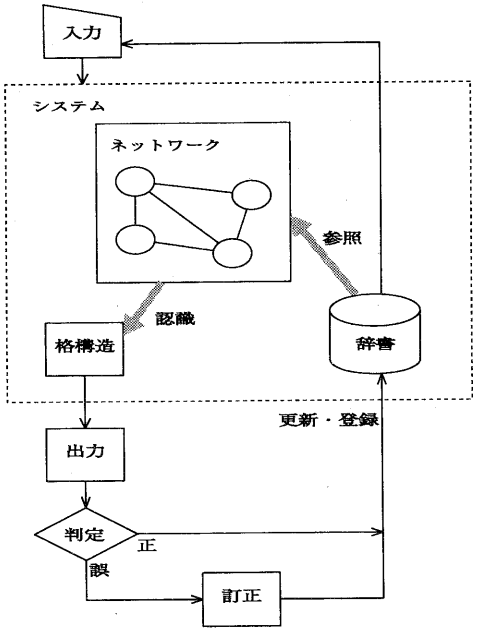

図 1: 処理の流れ

せ, 相互作用を起こす. 格構造は特殊な枝分かれするリンクで表現され, すべてのパーツの活性化状態が変化しなくなったときに相互作用が収束したものとして, システムは最も活性化しているリンクが表現する格構造を与えられた文の格構造であると認識し, 出力する.

3.2 枝分かれするネットワークと格構造表現

3.2.1 ノードとリンクの記述について

本稿では, 「ノード A」を 'A' と記述し, 「ノード A とノード B を結ぶリンク」を '{A-B}' と記述する. また, 本手法ではネットワークを枝分かれさせるために「'A' と '{B-C}' を結ぶリンク」といったリンクも用いるが, 同様に '{A-{B-C}}' と記述する.

ノード名には単語ノードには単語の表記をそのまま用いる. 概念ノードのノード名は '1.316' といった数字によるものと 'object' といった英語によるものがある. 前者はシソーラス [7] による分類番号を直接用いており, 後者は格概念を表現するものとして

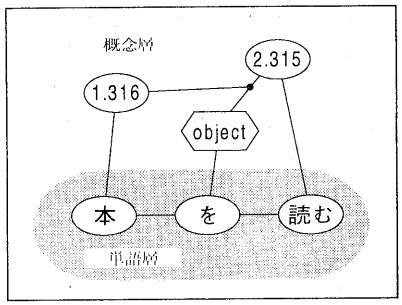

図 2: 格構造表現

本手法において導入したものである[1].

3.2.2 格構造表現

格構造は概念ノード間に張られたリンクによって表現される. ある動詞概念ノード 'V' と格概念ノード 'C' を結ぶリンク '{V-C}' によって動詞概念 V が格 C を取れることを表現し, そのリンク '{V-C}' への枝リンクとして名詞概念ノード 'N' が結合することによって格構造を表現する.

「本を読む」という文を例としてあげると, これを本手法のネットワークとして表現したものは図 2 のようになる. ここで, '本' の概念表現 '1.316' は, 'を' の概念表現 'object' や '読む' の概念表現 '2.316' とは直接結合せず, 'object' と '2.315' を結ぶリンク '{object-2.315}' と結合している.

このようなネットワークの形状が格構造を直接表現している. 図 2の例では '{1.316-{object-2.315}}' によって '2.315' の 'object' が '1.316' であることを表現され, さらに, '{ 本-1.316}','{ を-object}', '{ 読む-2.315}' といった単語ノードと概念ノードとを結ぶリンクにより, '読む' の対象格が '本' であることを表現しており, 「本を読む」の意味を表現している.

3.2.3 枝分かれするリンクの意味

本手法では, 助詞は単語同士を結合する機能的な働きをするものとしてではなく, 他の語と同様にそれ単独で何らかの概念を持つものとして考えている. 概念層では, ある 1 つのまとまった概念に 1 つのノー

[1]本手法では, 格概念ノードとして Fillmore の深層格に対応するもの 11 種類を用いている [2][3]

ドを与え, そのため, 格助詞の概念を表現するノードとして格概念ノードを設定する. 関連性の強いノード間にはリンクが張られるが, これは, その両端のノードの両方の概念を表現する1種のパイパーノードとして扱われる. 近い概念を表現するノード間に張られたリンクはこれによって大きな意味は持たないが, 異種の概念を表現するノード間にあるリンクは様々な意味を表現し得る.

本手法では, 動詞概念ノードと格概念ノードとの間のリンクによってある格を取り得る動詞を表現させ, また, それと名詞概念との間のリンクによって動詞と名詞の格関係を表現させているのみである. しかし, この手法ははかなり柔軟な表現が可能である. 例えば, 連体・連用修飾の概念を表現するノードを導入すると, 修飾の表現も可能であり, 順接・逆接の概念を表現するノードを導入すると, 2つ以上の文をその関係によって結び付ける表現も可能である.

3.3 活性化の相互作用

各パーツは互いに隣接しているパーツと相互作用する. つまり, 図2において'本'は'{ 本-を }', '{ 本-1.316}' と相互作用し, '1.316' は'{ 本-1.316}', '{1.316-{object-2.315}}' と相互作用する. リンクとノードは同様に取り扱われ同じ式によって活性値が計算される.

時刻 t におけるパーツ x の活性値 $V_x(t)$ は,

$$V_x(t) \quad = \quad V_x(t-1) + \sum_i \omega_{x\hat{x}_i} V_{\hat{x}_i}(t-1)$$

で表現される. ここで, $\omega_{x\hat{x}_i}$ はパーツ x とパーツ $\hat{x}_i$ との間の重み係数であり, $\hat{x}_i$ は x に隣接するパーツを表現している. 以下で $\omega_{x\hat{x}_i}$ について説明する.

3.3.1 ポジティブリンク

ポジティブリンクでは重み係数 $\omega_{x\hat{x}_i}$ は,

$$\omega_{x\hat{x}_i} \quad = \quad W_{x\hat{x}_i}$$

で計算される. 但し,

$$W_{AB} \quad = \quad C_\omega \sqrt{\frac{U_{AB}}{\sum_i U_{A\hat{A}_i}} \cdot \frac{U_{AB}}{\sum_j U_{B\hat{B}_j}}}$$

で与えられる. ここで, U_{AB} はパーツ A と B の接続頻度であり, 正しい解析結果で A, B が接続されているとき1増加する. この頻度情報の使用により. 使用頻度の高いパーツ同士がより強く相互作用する. また, C_ω は定数であり, 現在のところ, $C_\omega = 1$ として計算している.

W_{AB} はあるノードからそのノードに隣接するノードへの頻度と, それとは逆方向の頻度の相乗平均を取ることにより, 隣接しているパーツの多少によって活性値に差が現われないようになっている. 活性化はシステムに与えられた文によって活性化させられた単語ノードからの距離[2]が近いほど強くなる. また, 強く活性化しているパーツと多く隣接しているパーツも強く活性化する.

3.3.2 ネガティブリンク

相互作用を無駄なく高速に収束させるためにネガティブリンクを導入する. ネガティブリンクは, 1つの格概念ノードに複数のリンクが隣接しているとき, また, 1つの格概念ノードと他の概念ノードを結ぶリンク '{(格概念)-(概念)}' に複数のリンクが隣接しているとき, システムによってそれらのリンク間に張られる (図3). このとき重み係数 $\omega_{x\hat{x}_i}$ は以下のように計算される.

$$\omega_{x\hat{x}_i} \quad = \quad \begin{cases} W_{x\hat{x}_i} & (x \text{ がネガテイブリンク}) \\ -W_{x\hat{x}_i} & (V_x(t-1) < V_{\hat{x}_i}(t-1)) \\ 0 & (V_x(t-1) > V_{\hat{x}_i}(t-1)) \end{cases}$$

ここで, $\hat{x}_i$ は, $\hat{x}$ に隣接しているパーツのうちで x でないものを表わす. ネガティブリンクの計算にはあるリンクとその両端のパーツという3つのパーツが関係し, ネガティブリンクの両端のパーツのうち活性度が低い方の活性化値を抑制する.

3.4 拘束条件

活性値を発散させず, 確実に相互作用を収束させるために, 以下の2つのパラメータを設定し, 活性値を拘束する条件とする.

3.4.1 活性値総和

単語層, 概念層の各層において活性値の総和を一定とする. つまり各層において,

$$\sum_i V_{x_i}(t) \quad = \quad C_t \text{ (定数)}$$

[2]あるパーツとあるパーツへたどりつくために通らなければならないパーツの数が距離であり, 新しくリンクを登録することによりそれによって結合されるパーツ同士の距離が短くなる.

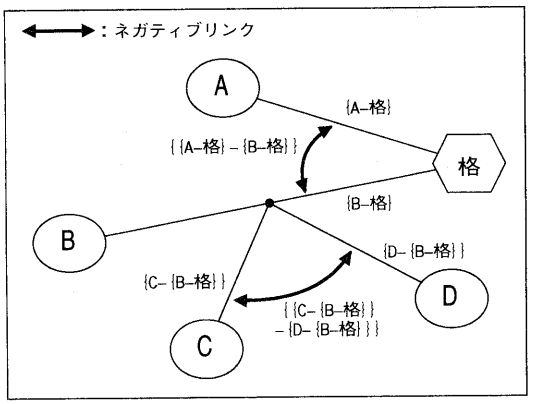

図 3: ネガティブリンク

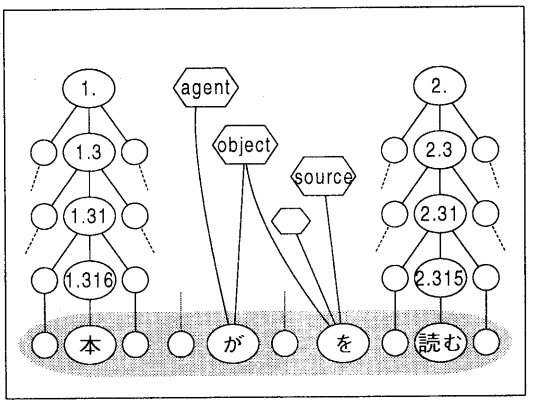

図 4: 辞書の構造

となる．ここで x_i は各層のパーツを表す．活性値の総和がこの値を超えたときには，その値が C_t となるように各ノードの活性値を正規化する．また実験では $C_t = 1$ としている．この拘束条件によって，ネットワークの一部の活性度が高くなると他の部分の活性度は相対的に低くなり，活性化の発散を抑制することができる．

3.4.2 活性値下限

パーツの活性値の下限 c_b を設定し，それ以下の活性値を 0 とみなす．この値と拘束条件により，ネットワーク上で活性化するパーツ数をある程度制限することができる．実験では拘束条件を固定し，活性値下限を $c_b = 0.01, 0.02, 0.05, 0.001$ と変化させその影響についても考察する．

3.4.3 相互作用の収束

すべてのパーツの活性値に変化がなくなったとき，相互作用が収束したものとして，そのネットワークが表現する意味の認識へシステムの作業が移る．

3.5 格構造の認識

活性化の相互作用が収束したとき，システムはネットワークの状態から格構造を認識する．システムは，それぞれの単語ノードから最も活性値の高いパーツを探索していき，格概念ノードに到達すると，元の単語ノードを格助詞と認識する．同様にして，格概念ノードと他のノートとの間のリンクに到達した場合には元のノードを動詞と認識し，他のリンクと接続しているリンクに到達した場合には名詞として認識し，さらに探索し，到達した格助詞をその名詞の格とする．

3.6 辞書の構造

単語ノードと概念ノードの接続情報は予め辞書に登録されている．以下のような手順で辞書を作成した．

- シソーラス [7] を参照し，掲載されている語を単語ノードとして，また，分類番号を概念ノード名として辞書に登録．対応する単語ノードと概念ノード間のリンクを登録．
- 格助詞を単語ノードとして，また，それぞれの格助詞が表現し得る格を格概念ノードとして登録．対応するノード間のリンクを登録．
- 格関係を表現するためのリンクは最初は登録されておらず，1 文解析毎にその結果を表現するリンクを頻度 1 として登録する．その際ユーザに問い合わせ，解析結果が誤っている場合には訂正してから登録する．また，すでに登録済みの場合には頻度をひとつ増加させる．

辞書の状態は図 4のようになり，これらの概念ノード間に格構造を表現するためのリンクが新たに登録される．

4 実験

以上のシステムをワークステーション上に作成し，実際に文の解析実験を行った．以下に，その手順と結果を示す．

4.1 手順

4.1.1 入力文

解析のための例文として，外国人のための日本語のテキスト [8] より単文 180 文 (格関係:300 組) を使用した．その単語数は 3 から 7 単語程度であり，1 文の平均単語数は 4.3 単語程度であった．システムは現在のところ修飾語の処理に対応していないため，文中に修飾語等が含まれる場合にはそれらを削除して用いた．また，これらの文は単語分けした状態でシステムに入力する．

4.1.2 出力・判定

システムは与えられた文の動詞と，その動詞に対する名詞の格を出力する．その出力に対してユーザが判定を行い，入力文中で確かにその格として使用されていると考えられる場合に正解析であるとする．それ以外は誤解析であるとし，ユーザに正しい解析結果を求める．

4.1.3 辞書

辞書はシソーラス [7] を用いて 3.6節で説明した手順で作成した．実験開始時には格関係を表現するためのリンクは登録されていない．1 文解析毎に与えられた文に対する格構造を表現するリンクが頻度 1 として登録される．また，その格構造を表現するリンクがすでに登録されている場合にはその使用頻度を 1 増加させる．

4.1.4 パラメータ

活性値の拘束条件を表すパラメータは，活性値総和を $C_t = 1$ として固定し，活性値下限を $C_b = 0.001, 0.002, 0.005, 0.01$ と変化させその影響について調べた．

4.2 解析結果

入力文と正解率の推移を表 1, 図 5に示す．解析文数が増加するに従い解析率も上昇し，文の登録と頻度情報導入の効果が見られる．$C_b = 0.001$ では正解率が最大 57.6%となった．

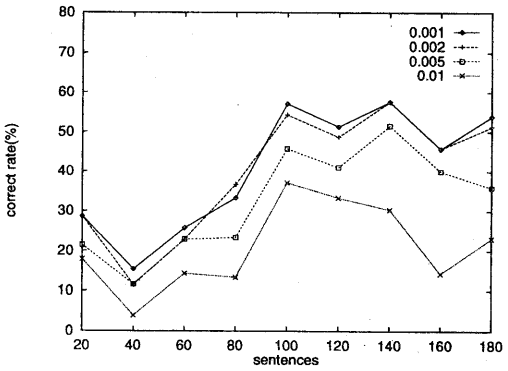

図 5: 実験結果

4.3 考察

4.3.1 解析率について

解析率については 50%程度であって高いとはいえない．しかし，本手法では解析のために単語同士の意味的な関連性しか用いておらず，構文情報などは使用していない．入力文が構文解析を行うことが可能なものであれば，構文解析を行い，その結果をネットワークの形状に反映させることが出来れば，より高い解析率を得られると考えられ，これは今後の課題である．また，本手法では解析に構文構造を用いていないために，構文解析が不可能な文であっても格構造解析が可能であり，非文法的な文を含む会話文の解析などに使用できるのではないかと考えている．

4.3.2 解析文について

本実験では表 1のような結果となったが，この段階では結果が一定の値へと収束しているとは言えず，更なる実験と考察が必要である．しかし，辞書への格構造の登録数が 100～200 程度という少ない段階から解析を行うことが出来るとも言うことができる．

表 1: 実験結果

活性値下限			0.001		0.002		0.005		0.01	
文数	格関係	平均単語数	正解	正解率 (%)	正解	正解率	正解	正解率	正解	正解率
1～20	28	3.8	8	28.6	8	28.6	6	21.4	5	17.9
21～40	26	3.6	4	15.4	3	11.5	3	11.5	1	3.8
41～60	35	4.5	9	25.7	8	22.9	8	22.9	5	14.3
61～80	30	4.0	10	33.3	11	36.7	7	23.3	4	13.3
81～100	35	4.5	20	57.1	19	54.3	16	45.7	13	37.1
101～120	39	4.9	20	51.3	19	48.7	16	41.0	13	33.3
121～140	33	4.3	19	57.6	19	57.6	17	51.5	10	30.3
141～160	35	4.5	16	45.7	16	45.7	14	40.0	5	14.3
161～180	39	4.9	21	53.9	20	51.3	14	35.9	9	23.1
合計	300	4.3	127	42.3	123	41.0	101	33.7	65	21.7

今回の実験で使った文は外国人のための日本語の教材のもので, 比較的簡単な文が多数であった. 今後はより複雑な文についての解析実験も行ないたいと考えている. そのためにはシステムに修飾表現などを解析するための拡張を施さなければならない.

4.3.3 誤解析について

誤解析の原因としては, 実験開始してから最初の段階は入力文の格関係を表現するリンクが登録されていないというものが多い. これはより多くの格関係を辞書に登録することによって回避できる.

入力文の格関係を表現するリンクが登録されている場合では, 解析に構文情報を使用していないことにより, 構文的には関係のない名詞概念ノードと格概念ノードが相互作用を起こしてしまうということが誤解析の原因へと繋がっている. これを回避するためには, 解析のために構文構造を使用する必要がある. 入力文が構文解析可能ならば, 構文解析を行ない, 係り関係に従って単語ノード間のリンクの接続を設定するといった方法が考えられる. 構文解析が不可能な文ならば, 単語間の意味関連性を用いて格構造解析を行なう本手法は有用である. そのような文についての頻度情報を導入していない状態での実験については [1] に示す. 頻度情報を導入した状態での実験と考察については今後の課題とする必要がある.

現段階ではシソーラスの分類番号の小数第 3 位までで概念の分類を行っているため, 1 つの概念ノードに接続している単語ノードの数が多数存在する. 1 つの入力文中に同じ概念ノードに接続する単語が存在した場合には 2 つ以上の単語が同じ概念として認識されてしまい, 誤解析の原因となっている. この回避のためにはシソーラスによる分類をより細かい数値まで使用するということが考えられる. また, 辞書を作成する際にシソーラスを用いるが, そのシソーラスによる単語の分類によって単語同士の意味距離が決定されるため, シソーラスの内容についても検討する必要があると考えられる.

解析に使用した入力文中では, 例えば, 格助詞「を」は 180 文中 80 回出現した中で, 72 回 (90%) は対象格を導くものとして扱われている. このような場合, 頻度の情報を導入したために, 「を」がそれ以外の格を導く文が入力されても, 単語ノード 'を' が概念ノード 'object' を活性化する割合が高くなり, 誤解析へと繋がる例も見られる. このように解析結果が例文に左右ため, 片寄りのない例文を使用する必要がある.

4.3.4 パラメータについて

活性値の下限を低く設定することによって多くのパーツが活性化し, 正解析を表現するリンクも活性化しやすくなり, 正解率を高めることが出来る. しかし, 余分なパーツが多数活性化することに起因する誤解析も少数であるが発生している. また多数のパーツが活性化すると処理が多くなるため解析時間が長くなるという問題も発生する.

また, 今回の実験ではパーツ間の重み係数を計算する際に使用する C_ω を常に $C_w = 1$ としているが, このパラメータの変化による解析率の変化も実験, 考察する必要がある.

5 おわりに

本報告では既提案システムである枝分かれするネットワークを用いた格構造解析システムへリンクの頻度情報を導入し, そのシステムの詳細について紹介した. また, その手法を用いて与えられた文の格構造解析を行うシステムを実際に作成し, 文の格構造解析実験を行い, その結果を示し, 考察した.

本手法では, 文の構文構造を用いずに格構造解析を行うことが出来る. また, そのためにネットワークを用いる. 日本語では格構造を表現するためには名詞, 動詞, 格助詞の 3 単語が必要であり, その格構造をネットワークによって表現するために, ネットワークを枝分かれさせ, 3 つの単語の概念に対応するノードを直接リンクで結び付けている. この特殊な形状のリンクは文の構造を柔軟に表現することが出来, 様々な文の表現に拡張することが出来ると考えられる. システムに文が与えられると, ネットワーク上の与えられた文中の単語対応する単語ノードが活性化され, ノード間に活性化の相互作用が起こる. その相互作用が収束した際のネットワークの活性化の状態の中で, 格構造を表現するリンクのうちもっとも強く活性化しているものが与えられた文の格構造を表現するものであるとしている.

そして, このシステムへある名詞がある動詞の何格になりやすいといった頻度情報を導入し, 180 文について解析実験を行い, 1 文解析毎に頻度情報を更新することにより, 文の解析率に向上を見ることが出来, また, 最大で 57.6%程度の解析率が得られ, 本手法が有用であったということが出来る.

本手法にはいまだシソーラスによる単語の分類や概念ノードの設定に関する問題, 格構造解析可能な文に対する格構造の扱いなどについて問題があり, 今度, 検討と改善が必要であると考えている.

謝辞

本研究の一部は科学研究費補助金 (課題番号 10680367) により行なわれた.

参考文献

[1] 竹内俊行, 荒木健治, 栃内香次: “枝分かれするネットワークを用いた格構造解析手法の性能評価”, 信学技報,NLC99-3,15–22(1999-5)

[2] 石崎俊:“自然言語処理”, 昭晃堂 (1995)

[3] 田中穂積, 辻井潤一:“自然言語理解”, オーム社 (1988)

[4] McClelland, J.L. and Rumelhart, D.E.: An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*,88,375–407(1981)

[5] Rumelhart, D.E. and McClelland, J.L. : An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*,89,60–94(1982)

[6] Waltz, D.L. and Pollack, J.B.: Massively Parallel Parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*,9,51–74(1985)

[7] 国立国語研究所:“分類語彙表”, 秀英出版 (1964)

[8] 益岡隆志, 田窪行則: “日本語文法 セルフマスターシリーズ 3 格助詞”, くろしお出版 (1987)