自然言語処理 120－20
(1997. 7. 25)

意味ネットワークからの文章生成

尾崎 正太郎　黒橋 禎夫　長尾 真
京都大学大学院 工学研究科 電子通信工学専攻
〒606-01 京都市左京区吉田本町
{s-ozaki, kuro, nagao}@kuee.kyoto-u.ac.jp

あらまし

従来の文生成で用いられてきた発話内容の表現は文や節の単位であった。しかし、発話内容は一般に漠然としていて文や節の単位に知識が整理されている場合は少ない。そこで、本稿では漠然とした発話内容を表現するために意味ネットワークを利用し、そこから文章を生成する枠組を示す。さらに、述語の選択において評価関数を用いることにより主題の変更が生じる際につながりがよくなることを示す。この枠組が意味ネットワークから文章生成の基礎となると考える。

キーワード　文章生成, 意味ネットワーク

Text Generation Using Semantic Network

OZAKI Syoutarou,　KUROHASHI Sadao,　NAGAO Makoto
Department of Electronics and Communication,
Faculty of Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-01, Japan
{s-ozaki, kuro, nagao}@kuee.kyoto-u.ac.jp

Abstract

In conventional text generation, sentences or clauses have been used as units for representing contents. However, the contents are generally vague and rarely arranged to form syntactically correct sentences or clauses. This paper presents a framework to generate texts using semantic network to represent the vague contents. Furthermore, a evaluation function for predicate selection is introduced here in order to improve the subject coherence. We believe this framework provides the basis for generating texts using semantic network.

key words　text generation, semantic network

1 はじめに

文の生成は早くから種々のアプリケーションシステムで実用的に用いられてきた。これは少しくらい不自然な文であっても利用者はその内容を理解できたからである。しかし、計算機で扱える知識・情報の増大にともない、人間が読んでも不自然さを感じさせないような良い文を生成することに対する期待が高まっている。

出力する情報量が少ない場合や前後のつながりを考えなくてもよい場合は、文単位で生成を行ってもかまわない。しかし、出力する情報量が増加すると文章として生成する必要がある。文章を自然なものにするには、文章を構成する文の間につながりが必要である。したがって、文章を生成するには、文単位の生成による出力を並べるだけでは不十分である。意味的な文のつながりを考慮した研究にはRST[1]などによるものがある。一方、表層的な文のつながりを考慮した研究には、主題の省略や、接続表現を扱ったものなどがある[2][3]。いずれの研究でも、生成すべき内容は、一般に、文または節の単位で用意されている。しかし、発話内容は一般に漠然としているので、文または節の単位で知識が整理されている場合は少ない。このため、本稿ではこの漠然とした発話内容を表現するために意味ネットワークを利用する。すなわち、発話内容に関連する概念を意味ネットワークの各ノードに割り当て、概念間の関係をリンクで表現する。そして、このネットワークからつながりのよい文章を生成することをめざす。

2 意味ネットワーク

2.1 意味ネットワーク

生成する文章のもとになるデータを、図1に示すようなネットワークの形で与える。

各ノードはそれぞれ概念を表す。概念は、事物を表すものと動作・様態を表すものに大別できる。図1では「チューリップ」「花」といった◯で示すノードが事物を表し、「咲く」「美しい」といった◯で示すノードが動作・様態を表している。ここでは、事物を表すノードを**体言ノード**、動作・様態を表すノードを**用言ノード**と呼ぶことにする。

リンクは、概念の間の関係を表す。接続しているノードの種類によって次のように分類される。

体言-用言リンク

動作・様態(用言ノード)には、それに対してある種の意味役割をはたす事物(体言ノード)が存在する。例えば、「咲く」には「いつ」、「どこで」、「何が」などの意味役割をもった事物が存在する。意味ネットワークでは、このような意味役割の関係を用言ノードと体言ノードを接続する体言-用言リンクとして表現する。

体言-用言リンクが表わす意味役割には以下のようなものがある[4]。

- agent：動作を引き起こす主体
  例)「花」-「咲く」
- object：動作・変化の影響を受ける対象
  例)「球根」-「植える」
- a-object：属性をもつ対象
  例)「花」-「美しい」
- time：事象の起こる時間
  例)「春」-「咲く」
- place：事象の成立する場所
  例)「北西部」-「盛んだ」

このような体言-用言リンクについては、用言ノードを親ノード、体言ノードを子ノードとして扱う。

体言-体言リンク

事物(体言ノード)間の関係としては次の2種類のものを考える。

1つは非常に限定的な関係で、たとえば「チューリップの花」という場合の「チューリップ」と「花」の関係である。この場合は「チューリップの花」を1つの概念と考え、意味ネットワークの1つのノードとして扱うことも考えられる。しかし、文章生成という立場では、「チューリップ」と「花」を分離して扱うことにより、より柔軟な文章生成が可能となる。そこで本稿では、このような問題を2つの体言ノードが**限定リンク**という特別のリンクで接続されているという状態で扱うことにする。限定リンクでは、限定されるノードを親ノード(上の例の場合「花」)、限定するノードを子ノード(「チューリップ」)とする。

もう1つは、上記の限定的な関係を除いた種々の意味関係を表わすリンクで、これを**属性リンク**と呼ぶことにする。簡単にいえば、属性リンクは「AはBだ」という形の名詞述語文として表現できるような「A」と「B」を接続するものである。このとき「A」がある属性を持ち、その属性値が「B」であるという関係にある。例えば次のようなものがある。

例)「チューリップ」-「多年草」
例)「オランダ」-「(チューリップの)産地」

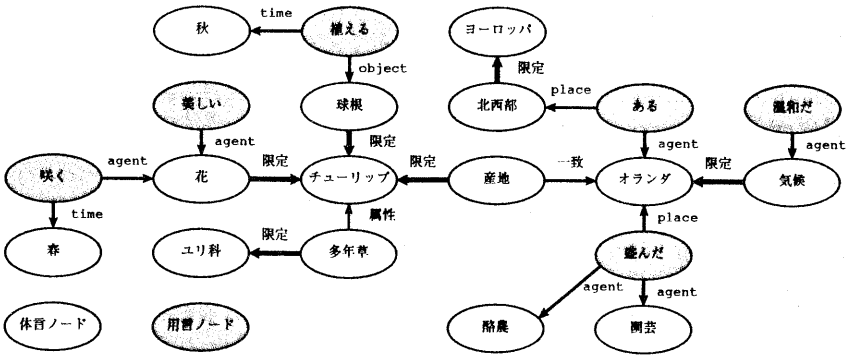

図 1: 意味ネットワーク

このように本稿では事物間の意味関係を「属性」という形で一般的にとらえている。「属性」にどのような種類がありえるかを明確に定義することは非常に難しい問題であるので、属性リンクをさらに細分類することは本稿では対象外とした。ただし、「オランダ」と「(チューリップの) 産地」との間の関係のように「A は B だ」としても「B は A だ」としても、それほど違いのない関係については「一致」として特別に扱うことにする。

なお、属性リンクについては属性値を表わすノードを親ノード、属性を持つ方のノードを子ノードとする。

用言-用言リンク

「みる」や「知る」といった動作・様態 (用言ノード) に対しては、「~するのを」といったように動作・様態が補語となることがある。このような用言ノードが用言ノードの補語となるものを用言-用言リンクとして表現する[1]。このような用言-用言リンクについては意味役割をはたす用言ノードを子ノード、他方のノードを親ノードとする。

まとめると、本稿で扱う意味ネットワークは、動作・様態を表わす用言ノード、事物を表わす体言ノード、体言-用言リンク、体言-体言リンク (限定リンクと属性リンク) および用言ノードを補語とするような用言-用言リンクからなるものである。

[1] その他に、動作・様態 (用言ノード) 間の意味関係としては、条件、因果関係、時間的順序など様々な関係が考えられる。しかし、本稿では扱わないことにする。

3 文章生成の手続き

3.1 文章の生成

意味ネットワークの形で表現された発話内容を入力とし、意味ネットワーク全体を表現するような文章を生成することを考える。文章は単に文が並んだものではなく文と文の間になんらかのつながりを必要とする。

ここでは、それぞれの文が主部 (主題とその修飾語) と述部 (述語とその修飾語) からなるものとして、文の間のつながりを考える。ただし、主題ノードは常に体言ノードとする。

文のつながりを感じさせるものに主題の連続によるものがある。

例) チューリップはユリ科の多年草です。
　　(チューリップは) 春に花が咲きます。

この例では主題の「チューリップ」が連続することにより文のつながりがえられている。このような文のつながりを**主題連続**と呼ぶことにする。

もう 1 つの文のつながりとして、前の文に現れた主題以外の語が次の文で主題になるというものがある。

例) チューリップの産地はオランダです。
　　オランダは酪農や園芸が盛んです。

このような文のつながりを**主題遷移**と呼ぶことにする。

図 1の意味ネットワークで主題連続を考慮せずに主題の選択をランダムに行うと、表 1の例 1 に示すように主題が頻繁に変わり文章の体をなさない。また、できるだけ主題を連続するようにした場合でも、主題遷移について考慮しないと文のつながりがわるくなる。例えば、表 1 の例 2 では、「チューリップ」が 2 文めに現れた後に他の話題に移り再び 5 文めで主題となるため、文のつながり

表 1: 出力順序の変更例

例 1	例 2	例 3
オランダは気候が温和です。 チューリップはユリ科の多年草です。 ヨーロッパは北西部にオランダがあります。 春は美しいチューリップの花が咲きます。 チューリップの球根は秋に植えます。 オランダはチューリップの産地です。 オランダは酪農や園芸が盛んです。	オランダは気候が温和です。 秋に球根を植えるチューリップの産地です。 ヨーロッパの北西部にあります。 そこでは酪農や園芸が盛んです。 チューリップは春に咲く花が美しいです。 ユリ科の多年草です。	オランダは気候が温和で、酪農や園芸が盛んです。 ヨーロッパの北西部にあります。 花が美しいチューリップの産地です。 チューリップはユリ科の多年草で、春に花が咲きます。 秋に球根を植えます。

がわるい。それに対して例 3 では、全ての文が主題連続または主題遷移のいずれかのつながりをもち、文のつながりがよい。

そこで、以下では主題連続・主題遷移による文のつながりを考慮しながら、文章生成において主題・述部・表層表現のそれぞれを決定する方法について述べる。

3.2 主題の決定

ある主題について出力すべき内容が残っている状態でほかの主題に移ると、もう 1 度もとの主題にもどる必要がある。このようにして主題が頻繁に変わると文のつながりがよくないので、ある主題について出力すべき述部があるかぎりその主題を連続する。これにより主題連続による文のつながりがえられる。

前文の主題で出力すべき述部がない場合は、前文の述部に含まれるノードを主題とすることを試みる。主題となりうるノードが複数ある場合は任意のものを選択する[2]。

直前の文に含まれるノードを主題として文を生成できない場合には、さらに前の文に現れたノードを主題とする必要がある。文のつながりをよくするためには、できるだけ近い位置にある語を主題としたほうがよい。そこで、今までに出力したノードをスタックを用いて保存し、できるだけ近い位置にある語が主題となるようにする。

3.3 述部の決定

ある主題に対して述部になりうるものが複数得られることがある。この場合、述語をどのような順序で出力するかによって文のつながりに違いが生じる。

例えば、図 2のようにある主題に対して出力すべき述部に述部 1 と述部 2 の 2 つがあり、述部 2 にはさらに出力すべき述部 3 が接続している場合を考える。この場合の出力の順序としては

a. 述部 1 $\overset{\text{主題連続}}{\rightarrow}$ 述部 2 $\overset{\text{主題遷移}}{\rightarrow}$ 述部 3

b. 述部 2 $\overset{\text{主題連続}}{\rightarrow}$ 述部 1 → 述部 3

c. 述部 2 $\overset{\text{主題遷移}}{\rightarrow}$ 述部 3 → 述部 1

の 3 通りが考えられるが、これらのうち、それぞれの文において主題連続・主題遷移によるつながりがえられるのは a のみである。つまり、この例では述部 1 を先に出力するほうが文のつながりがよくなる。

以上の議論は、意味ネットワークのどれだけの範囲を 1 文とするかに関係する。例えば、図 3のような場合、それぞれの文を単文とすると述部 1 と述部 3 のいずれを選んでも主題遷移によらない主題の変更を生じるが、連体節[3] による複文で出力することにすれば

- 述部 1,2 $\overset{\text{主題連続}}{\rightarrow}$ 述部 3,4 $\overset{\text{主題遷移}}{\rightarrow}$ 述部 5

のように出力できるので、主題連続・主題遷移による文のつながりがえられる。ただし、1 文に含まれる連体節の数が多いと文の構造が複雑となり文が理解しづらくなるので、何らかの制限を加える必要がある。ここでは、1 文に含まれる連体節の数の上限を 1 とすることで連体節に対する制限とする。

これまで述べてきたように、出力すべき述部が複数ある場合には話題が展開する述部を後に出力したほうがよい。実際に述部から話題が展開する様子を知るには文章生成で用いるノードをすべて取り出す必要があるが、これは困難である。

そこで、述部から話題が展開する度合いをそれぞれの述部に対してつくることのできる連体節の数を調べるこ

[2] この場合、どのような順序で主題を選択しても必ず主題遷移によらない主題変更を生じるので、主題の変更という点からの文のつながりに違いを生じない。

[3] 連体節には限定的な修飾を表す場合と非限定的な修飾を表す場合があるが、本稿では扱わないことにする。

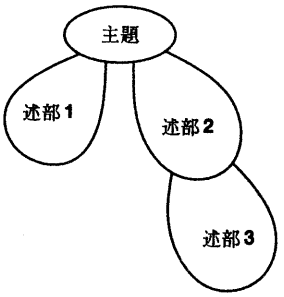

図 2: 述部の選択の例 1

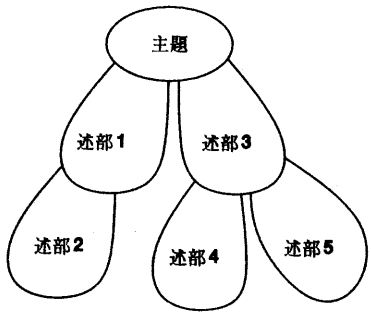

図 3: 述部の選択の例 2

とによって見積もり、連体節の少ない述部から先に出力することにする。

また、述部に対してつくることのできる連体節が複数ある場合には、連体節が長くなると文の構造が複雑になって読みにくい文になることを考慮して候補の中でノード数の少ないものを選択することにする。

連体節は主題に対してつくることもできるが、新しい情報が主題に対する連体節として表現されるのは不自然なので、主題に対して連体節となりうる述部は独立した文または次に述べる並列節として出力することにする。

短い文が連続すると断片的な印象を与えるため、並列節を用いて表現したほうがよい場合がある。例えば、「オランダはチューリップの産地です。気候が温和です。」とするよりは「オランダはチューリップの産地で、気候が温和です。」としたほうがまとまりが感じられる。そこで、連体節をもたない述部については、述部のノード数の少ないものから順に出力し、述部のノード数がしきい値以下であれば、別の述部に対する並列節として出力することにした。ここでは述部のノード数のしきい値を 2 とした。

これまで述べた手続きの例を図 1の意味ネットワークを用いて示す。最初の主題ノードを「オランダ」とすると「気候が温和だ」「酪農や園芸が盛んだ」「ヨーロッパの北西部にある」「チューリップの産地だ」の 4 つが述部となりうる。それぞれの述部に対してつくることのできる連体節の数は「チューリップの産地だ」のみが 3 で他は 0 であるため、「チューリップの産地だ」以外の述部から出力することになる。まず、ノード数の最も少ない述部「気候が温和だ」が選択される。述部「気候が温和だ」のノード数は 2 で並列のしきい値以下となるため「酪農や園芸が盛んだ」と並列節をつくる。つぎに述部「ヨーロッパの北西部にある」、述部「チューリップの産地だ」の順で出力される。「オランダ」を主題として出力すべき述部がなくなったので、前文の述部中の「チューリップ」を新たな主題として残りの文章を生成する。

3.4 表層表現への変換

取り出されたノード群を文の形にするには、これらを文字の並びに変換する必要がある。この過程は、ノードおよびリンクに表記を対応させ、それを適切な順序で並べることにより行われる。ここでは、まずノードを出力する順序について述べ、次に語が述語となる場合の表記について述べる。

リンクとその両端のノードを表層表現に変換するには、ノードに対応する語の表記およびリンクに対応する表記を用いて

- < 子ノード >< リンク >< 親ノード >

の順序で並べる。

ノードに対して子ノードが複数ある場合には< 子ノード >< リンク >の部分が複数になり、それらの順序を決める必要がある。一般に日本語文において述語を修飾する要素の並べ方には標準的な順序があるとされている。例えば、主体を表す要素は時間や場所を表す要素の後に置かれることが多い。そこで、リンクの間に次に示すような順序をつけ文中では上位のものを前方に置くことによりノードを並べる順序を決定する。

1. time (事象の起こる時間)
2. place (事象の成立する場所)
3. agent (動作を引き起こす主体)
4. object (動作・変化の影響を受ける対象)

ある種のリンクは出力の順序によって異なった表記をとる。例えば、「A が B から離れている」と「B が A から離れている」では「離れている」に対して「A」と「B」は

同じ関係である。このようなものについては特別にルールを記述しておく。

用言-用言リンクの表記は、親ノードによって「できる」のように「基本形+ことが」が適切なものや「見る」のように「基本形+のを」が適切なものがあるのでそれぞれ記述しておく。

以上で述べたようにしてノードとリンクを表層表現に変換することにより、主部と述部の表層表現が得られる。文の表層表現は、主部と述部のそれぞれの表層表現を

「＜主部＞は＜述部＞。」

のような順序で並べることにより得ることができる。主題を示すのに「は」以外の表記が用いられることもあるが、本稿では「は」以外の表記は扱わないことにする。

語が述語として働く場合には、接尾辞や判定詞を接続して次のような表記をとることにする。

- 動詞:基本連用形＋「ます」
- イ形容詞:基本形＋「です」
- ナ形容詞:語幹＋「です」
- 判定詞:「です」

語が述語として働く場合には、表現している動作・様態の置かれている時間的な位置づけあるいは継続や完了といった局面によって様々な形をとる。例えば、時間的な位置づけを表すのには「～する」「～した」などの表現が用いられる。また、継続や完了といった局面を表すのには「～している」「～してしまう」などの表現が用いられる。これらのテンス・アスペクトの情報は意味ネットワークの各ノードに対して与えておく。テンス表現としては基本形とタ形、アスペクトの表現としては「テ形+いる」が出力できるようにしている。

連体節や並列節と主節との間の接続表現は様々なものをとりうる[4]が、本稿ではその選択は扱わないことにし以下のような接続表現を用いることにした。

- 連体節
  - 動詞:基本形
    例)「咲く」
  - イ形容詞:基本形
    例)「美しい」
  - ナ形容詞:連体形
    例)「盛んな」
  - 体言＋判定詞:連体形
    例)「多年草の」
- 並列節
  - 動詞:基本連用形
    例)「咲き、」
  - イ形容詞:基本連用形
    例)「美しく、」
  - ナ形容詞:タ系連用形
    例)「盛んで、」
  - 体言＋判定詞:タ系連用形
    例)「多年草で、」

文中での連体節、並列節の配置については次のように行う。連体節は修飾する体言の前に置かれるので

- ＜連体節＞＜体言ノード＞

のような順序で出力する。限定リンクと連体節では限定リンクの方がノードとの結びつきが強いと考え、体言ノードに対して限定リンクによって接続しているノードと連体節とがある場合には

- ＜連体節＞＜子ノード＞＜リンク＞＜親ノード＞

の順に並べる。並列節は

- ＜並列節＞＜主節＞

の順序で出力する。

同一の主題が繰り返されると不自然となるため、同一の主題が連続するときは2回目以降の主題の省略を行う。主題の省略が続くと、内容がはっきりしなくなるため主題が表現されてからの省略の回数がしきい値を越えるたびに主題を以下のように表現する。

- 主題が場所であれば指示詞を用いて「そこ」と表現する。
- 主題が場所以外の場合は主題を繰り返す。これは「それは」とすると前文をうけることがあるからである。

しきい値は2とした。

4 生成例と考察

図1の意味ネットワークにおいて最初の主題を「オランダ」とすると表2の生成例1が得られる。主題「オランダ」が連続し、「チューリップ」への主題の変更は主題遷移のつながりがある。最初の主題を「チューリップ」とすると表2の生成例2が得られる。主題が「オランダ」に変更されるときに主題遷移のつながりがある。

図4の意味ネットワークにおいて最初の主題を「ごん」とすると表2の生成例3が得られる。5文めの主題「森」以外は主題連続または主題遷移によるつながりがある。最初の主題を「森」とすると表2の生成例4が得られる。5文めの主題「山」以外は主題連続または主題遷移によ

[4] 連体節では「～する」「～した」、並列節では「～し」「～して」「～したり」といった表現が可能である。

表 2: 生成例

生成例 1	生成例 2
オランダは気候が温和で、酪農や園芸が盛んです。 ヨーロッパの北西部にあります。 花が美しいチューリップの産地です。 チューリップはユリ科の多年草で、春に花が咲きます。 秋に球根を植えます。	チューリップはユリ科の多年草で、花が美しいです。 秋に球根を植えます。 春に花が咲きます。 チューリップの産地は気候が温和なオランダです。 オランダは酪農や園芸が盛んです。 ヨーロッパの北西部にあります。
生成例 3	**生成例 4**
「ごん」はひとりぼっちで、子どものきつねでした。 山にある森にすんでいました。 昼や夜に山に近い村でいたずらしました。 山は中山からすこしはなれていました。 森はしだがいっぱいしげっていました。	森はしだがいっぱいしげっていました。 村に近い山にありました。 そこにはひとりぼっちの「ごん」がすんでいました。 「ごん」は子どものきつねで、夜や昼に村でいたずらしました。 山は中山からすこしはなれていました。
生成例 5	**生成例 6**
WWW は「World Wide Web」の略で、ハイパーテキストを表示するブラウザでアクセスします。 ブラウザはテキストをもってくることができます。 ハイパーテキストはテキストへのポインタをもちます。	ブラウザは「World Wide Web」の略の WWW にアクセスします。 テキストへのポインタをもつハイパーテキストを表示します。 テキストをもってくることができます。

るつながりがある。

図 5の意味ネットワークにおいて最初の主題を「WWW」とすると表 2の生成例 5 が得られる。最初の主題を「ブラウザ」とすると表 2の生成例 6 が得られる。いずれの例も主題によるつながりがある程度えられている。

これらの生成例から考察すると今後の課題として以下のようなことがあげられる。

- ノード間の意味的な結びつき

  出力の順序を決定する際に意味的な結びつきも考慮したほうが自然な文となる。例えば、生成例 1 の 1 文めの「園芸」と 3 文めの「チューリップ」は意味的なつながりがあるので文章中で近い位置関係になるように出力したほうが自然である。

- 並列の類似度

  並列節をつくる際に節の間の類似度が高いものは優先的に並列としたほうが自然な文となる。例えば、生成例 1 の 4 文めの「春に花が咲きます」と 5 文めの「秋に球根を植えます」は、「春」と「秋」、「花」と「球根」「咲く」と「植える」の対応があり並列となるほうが自然である。

- 連体修飾節

  今回は、連体修飾節の制限を数によって行ったが節の長さなどの要因も考慮する必要がある。また、連体修飾節をつくった場合に限定用法で働くか非限定用法で働くかを区別する必要がある。

- 述部の内容

  本稿では述部の内容を出力順序や表層表現に反映させていない。しかし、述部の内容が出力順序に影響を与えるものとして述部のうち定義的なものを先に出力するといったことが考えられる。また、述部の内容が表層表現に影響を与えるものとしては、定義的な内容であれば「…は～。」といった表現よりも「…とは～である。」といった表現のほうが適切であるということがある。

- 時間的順序・因果関係

  今回は時間的順序や因果関係を対象外としたが、これらの関係を含む意味ネットワークから文章を生成する場合にそれらが出力の順序に与える影響については検討する必要がある。

- 巨大なネットワークの場合

  今回用いた意味ネットワークは主題の変更が 1 回程度で出力できるものであった。より大きなネットワーク

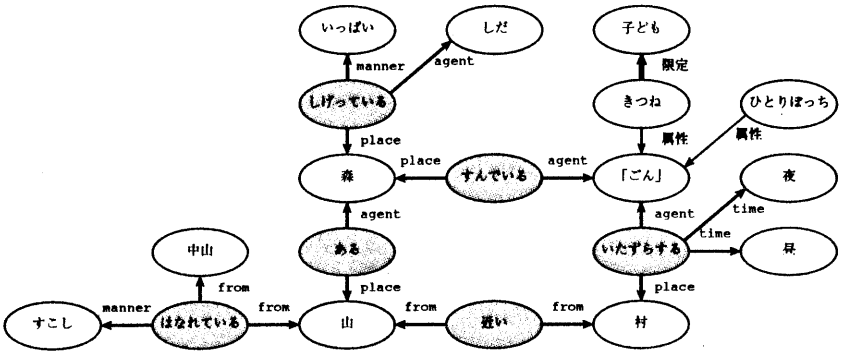

図 4: 意味ネットワーク

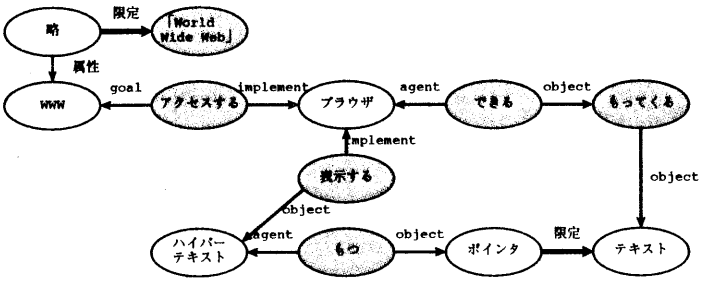

図 5: 意味ネットワーク

でノードの選択や主題の表現が適切に行えるかどうかを確認する必要がある。

5 結論

漠然とした発話内容を表現するために意味ネットワークを利用し、そこから文章を生成するために必要な枠組を提案した。述語の選択において、述部からの話題の展開を連体節の数によって見積もることにより文のつながりがある程度自然になることを示した。これらが意味ネットワークからの文章生成の基礎となると考える。

参考文献

(1) W.C.Mann and S.A.Tompson: "Rhetorical structure theory: Description and construction of test structures", Natural Language Generation (Ed. by G.Kempen), Martinus Nijhoff, chapter 7 (1987).

(2) 高橋, 桃内, 宮本："汎用文章生成システムによる日本語主題表現生成方略の実現", 情報処理学会 自然言語処理研究会 NL56-2 (1986).

(3) 高橋, 桃内, 宮本:"文章生成における接続詞の生成方略について", 情報処理学会自然言語処理研究会 NL62-3 (1987).

(4) 日本電子化辞書研究所："EDR コーパス 1.5 版" (1996).