自然言語処理 97−14
（1993. 9. 17）

語義曖昧性を考慮した有意な語彙連鎖の生成

本田 岳夫, 奥村 学
北陸先端科学技術大学院大学 情報科学研究科

[概要]

本稿では、語彙結束性を表す語彙連鎖をシソーラスに基づいて生成する手法について述べる。語彙連鎖はテキスト内の関連語の連鎖であり、テキストの文脈情報を与える。我々は、連鎖生成中の文脈を表現するため、疑似スタック構造を語彙連鎖生成過程に導入する。疑似スタックは、生成中の連鎖の順序を最近更新された連鎖が上位に来るように制御する。語はスタックの上位にある連鎖から順に結束性を調べることで、語の近傍の文脈を得ることができるので、疑似スタックを用いて語彙連鎖を漸進的に生成する過程で語義曖昧性を同時に解消することができる。これによって、意味のない連鎖が生成されないようにすることもできる。

また、意味のない不要な連鎖を取り除き、有意な連鎖を取り出すために、タイトル語情報と、一部の構文的情報を利用し、実際に連鎖の取り出しを行ない、その有効性を考察する。

Computing Meaningful Lexical Chains while Resolving Word Sense Ambiguity

HONDA Takeo,OKUMURA Manabu
Faculty of Information Science, Japan Advanced Institute of Science and Technology
(Tatsunokuchi Ishikawa 923-12 Japan)

Abstract

In this paper, we present a method of computing lexical chains by thesaural relations. A lexical chain is a chain of related words as an indicator of the text structure. When computing lexical chains, we introduce a pseudo-stack structure to represent contextual information. The pseude-stack structure impliments an order of lexical chain to put a recent updated chain on the top of the stack. Computing lexical cohesion between a word and a lexical chain on the stack from the top to the bottom, contextual information can be found. So using a pseudo-stack structure,we resolve the word sense ambiguity and get meaningful lexical chains incrementally.

Also,we present a naive criteria for removing meaningless lexical chains and getting meaningful lexical chains using information from heading information and syntactic information.

1 はじめに

テキストをテキストとして成り立たせている「つながり」には大きく分けて二つの側面がある。一つは、言語形式を拠り所とする結束性で、もう一つは、談話における意思伝達行為という形で作用する首尾一貫性である。本研究では、二つの「つながり」のうち結束性に注目し、その中でも、語の間の意味的な関係から得られる語彙結束性を扱う。

テキスト中で語彙結束性関係にある語のまとまりを、語彙連鎖と呼ぶ。語彙連鎖はドメインを限定しないテキストに存在するので、談話構造の解析のための指標として、また、語義曖昧性解消などの時の文脈情報として利用できる。本稿では疑似スタック構造を用いて、語彙連鎖を漸進的に生成する手法を示し、実際のテキストに適用して、自動的に語彙連鎖の生成を試みる。疑似スタック構造は、生成中の連鎖の順序を最近更新された連鎖が上位に来るように制御しているものである。スタック上の上位にある連鎖がその近傍の文脈を与えていると考えられるので、スタックの上位の連鎖から順に結束性を調べることで、語の近傍の文脈情報が得られるので、語彙連鎖を漸進的に生成する過程で語義曖昧性を同時に解消することができる。これによって、意味のない連鎖の生成を抑えることもできる。

しかし、こうして生成した語彙連鎖には、テキストの意味のまとまりを示していたり、テキストの構造を示した有意な語彙連鎖ばかりでなく、意味のない不要な連鎖も同時に生成されてしまう可能性がある。本研究では、生成した連鎖から有意な連鎖を取り出すために、タイトル語情報と、一部の構文的情報を利用し、実際に連鎖の取り出しを行ない、その有効性を考察する。

英語を対象に語彙連鎖を求める研究が Morris and Hirst によって行なわれている [1]。この文献ではシソーラスに *Roget's International Thesaurus* を用いて、語彙連鎖 を計算する手法を提案している。しかし、実際に計算機上で実現していない、連鎖を作るとき語義曖昧性を考慮していない、などの問題点がある。山本、増山、内藤は、手がかり語と語彙結束性 を併用した日本語テキストの段落分けの手法を提案している [5]。しかし、語義曖昧性を考慮していないため、意味のない語彙結束性関係が計算される可能性がある。佐々木、増山、内藤は、語義曖昧性解消をし、語彙結束性を結束チャートで表す手法を提案している [4]。しかし、出現する語すべてについて語彙結束性を計算しているため、生成されるチャートには不要な情報も含まれてしまう可能性がある。

本稿の構成を次に示す。第 2 節では、語彙結束性について説明する。第 3 節で語彙連鎖を生成する手法について述べ、つづく第 4 節で、実際のテキストに適用して語彙連鎖を作成し、考察を加える。第 5 節では、作成した連鎖から有意な連鎖を取り出す条件を示し、その有効性を考察する。最後に第 6 章で結論と今後の課題を述べる。

2 語彙結束性

複数の文からなるテキストに用いられている語の間には意味的な関係がある。この意味的関係のことを語彙結束性と呼ぶ。例えば、

example 1
ともかくこうして膨張を続ける宇宙の中で数多くの星が誕生、消滅を繰り返しました。そして宇宙の誕生から約１００億年後、他の星と同じ様にして、宇宙の一部の星間雲が万有引力により収縮し、原始太陽を中心にして原始太陽系星雲と呼ばれるガスの円盤を作りました。[1]

[1]柳哲雄,「海の科学」[8] より

という文章では意味的に関係がある語のあつまり、[星, 星, 星, 太陽, 太陽系, 星雲] がある。この語彙結束性関係のある語の集まりのことを語彙連鎖と呼ぶ。この語彙連鎖はシソーラス上の同一のカテゴリ (意味分類) に属するものとして、計算する。なお、本研究では語彙結束性を計算する尺度としてシソーラスに分類語彙表 [3] を用い、分類語彙表の同一のカテゴリに属するものが語彙結束性関係にあるとする。

連鎖を生成する時に問題になる点の一つに語義曖昧性の問題がある。語の意味は孤立して存在するわけではないので、語の意味を解釈するためには、文脈に関する知識が不可欠である。例えば、「地球」という語は二つの語義を持ち、一方は、「太陽」、「月」、「火星」などの語と同じカテゴリ (天体) に属すもので、他方は、「地殻」、「マントル」などの語と同じカテゴリ (地) に属すものである。ここで、[星雲、星、太陽系、惑星] という文脈が与えられれば、「地球」は「天体」という意味に限定できる。

また、**example 1** の中に現れる語で、「中」,「後」,「間」という語が語彙結束性があるとして、連鎖 [中, 後, 間] を作ったとしても、**example 1** の中ではこの連鎖に意味がない。有意な連鎖を生成するためには、[星, 星, 星, 太陽, 太陽系, 星雲] のような意味のある連鎖のみを取り出し、[中, 後, 間] のような意味のない連鎖を除かねばならない。

3 語彙連鎖の生成

テキスト中で、ある話題が進行している時には、その話題を表す語彙連鎖が現れ、話題が変わると別の連鎖が現れる。我々は、この現象をモデル化し、語義曖昧性解消をしながら語彙連鎖を生成する枠組として疑似スタック構造を導入した。この疑似スタック構造は、生成中の連鎖を最近更新された連鎖がスタックの上位にくるように連鎖の順序を制御したものである。つまり、ある時点でスタックの上位にある連鎖が、その近傍の文脈を表している。語を連鎖に追加する時にはスタックの上位にある連鎖から順に語彙結束性を調べ、最初に見つかった連鎖に語を追加する。多義語の場合には、最初に見つかった連鎖が近傍の文脈を与えてるのでその連鎖と結束性関係にある語義が多義語のその時点の語義になる。

疑似スタック構造では、その時の話題を表す連鎖がスタックの上位に存在し、話題が変われば別の連鎖が上位に現れるので、スタック中の連鎖の動きを観察することによって、談話構造の解析に役立てることができる。

3.1 アルゴリズム

本手法は漸進的に語彙連鎖を生成する。以下にアルゴリズムを示す。まず、テキストの形態素解析をする。(日本語形態素解析システム JUMAN[6] を利用した。) つぎに形態素解析の結果から、連鎖を構成する候補となる語を選択する。ここで選択される語は、名詞、動詞、形容詞に限り、さらに名詞は「こと」などの形式名詞、「(〜の) ため」などの副詞的名詞を除き、動詞は「〜によって」「〜について」などのように附属語的に使われているものを除いた。

つぎに選択した候補語に対して連鎖を生成する。まず、一文の中での候補語に対して、結束性関係を調べる。曖昧性のある語が、この時点でどれかの語と結束性関係にあるならば、その時の語義を採用することにする。これは文の内部での結束性関係の強度が文間の結束性関係の強度より強いであろうという推測に基づく。

一文内結束性関係を調べることで部分的に曖昧性解消をした候補語を、疑似スタック内の連鎖に追加する。まず、新たに連鎖に追加しようとしている語の語義と、スタック中の連鎖との結束性関係を調べ、どの連鎖とつながるか求める。このとき、スタックの上位につまれたものから順に調べ、結束性関係にある連鎖が見つかった時点で、その連鎖より下にある連鎖とは結束性関係を調べない。

次に 更新された連鎖に含まれる語に曖昧性があるかどうかを調べる。疑似スタックとは別に曖昧性のある語を記録しておき、更新された連鎖の中の語に曖昧性があれば、その連鎖とは別の連鎖に含まれる語を

連鎖から取り除く。

語をスタック上の連鎖と結束性関係を求める際、どの連鎖とも結束性関係にないものがある場合がある。その語は後に現れる語と連鎖を構成可能性があると考え、一語のみの新しい連鎖を作り、スタックの先頭に積む。このとき、候補となる語に曖昧性があるならば、曖昧性のある語として、記録に追加し、それぞれの語義に分解し、複数の連鎖を作る。

以上の操作を文章末まで繰り返す。

このアルゴリズムを **example 2** に適用した例を示す。

example2

..... a) これと同じようなことが原始地球でおこったのです。.....。b) そして..... マントルを形成し、..... ..... 岩石が地球表面に薄い地殻を形成したのです。.....。c) 地球の水分のほとんどは蒸発して、.....

「地球」には前出のように [天体][地] の 2 つの語義がある。文 a) の「地球」は、一文内の結束性およびそれまでの連鎖からでは語義曖昧性が解消されないため、スタックには二つの「地球」を含む連鎖がある。ここに「マントル」を追加する。スタックの上位にある連鎖から順に結束性を調べ、「マントル」は [地] の連鎖につながる (図 1 左上)。「マントル」を追加した連鎖をスタックの先頭に持っていき、この連鎖に曖昧性があるか調べる。ここでは「地球」に曖昧性があった (図 1 右上)。「地球」を他の連鎖から消去する (図 1 左下)。文 b) の「地球」は同一文に「マントル」「地殻」があるために [地] の語義に決定できるが、文 c) の「地球」は一文内では語義の決定ができず、曖昧性がある。スタックの上位にある連鎖から順に結束性を調べていくと、[地] の連鎖が上位にあり、文 c) の「地球」の語義が [地] に決定できる (図 1 右下)。

4 実験

前節のアルゴリズムを SICStus Prolog 上に実現し、実際のテキストを用いて実験した。テキストに「海の科学」[8] を使用した。このテキストは、12 の章からなり、各章はいくつかの節からなる。節はいくつかの形式段落からなる。実験にはこのうち第 1 章を用いた。

第 1 章には候補語 481 語中 166 語に語義曖昧性があり、平均語義数は 1.91 であった。一文内の結束性で解消できる曖昧性は 67 語 (40%)。最終的に曖昧性解消されなかった語は 7 語であった。

4.1 考察

最終的に曖昧性の解消されなかった語は、複数の語と結束性関係を持たなかった語、つまり、疑似スタック上の連鎖としては一語のみの連鎖になっているものである。曖昧性を解消できたもので、誤っているものは次のようなものがあった。

- 形態素解析誤りによるもの。例えば、「当初」という語が「当」「初」に分かれ、それぞれ別に連鎖を構成した。
- 地殻など「地」に関する「地球」の語義になるべき語が、それまでの文脈にひきずられて、「天体」の「地球」になってしまった。
- 「内外」「上下」「前後」などの関係を表す語に多く見られるが、同一文内にシソーラスの同一カテゴリに属する語があるために誤った連鎖が生成されてしまう例もある。

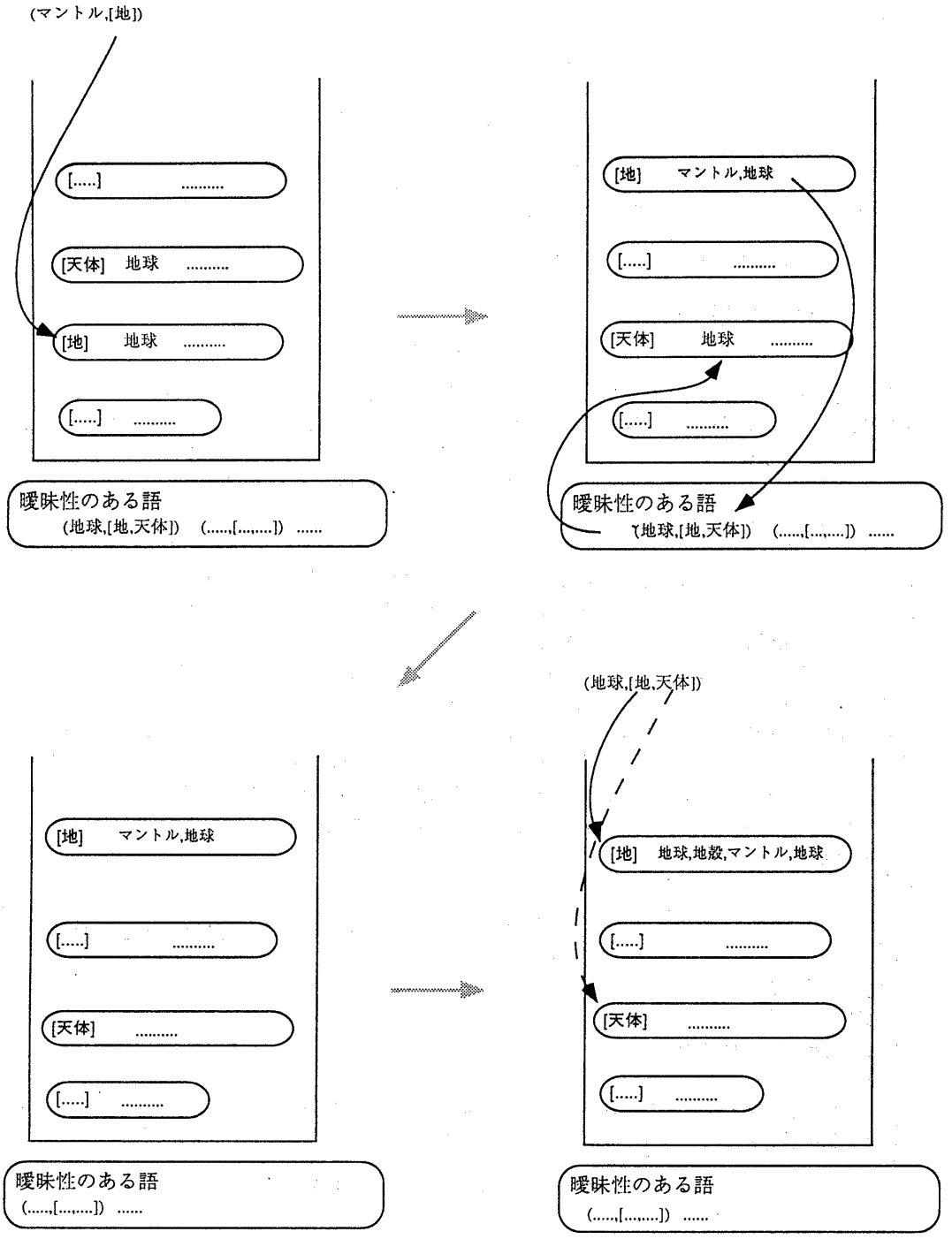

図 1: 疑似スタック上での連鎖生成の様子

5 有意な語彙連鎖

前節までで求めた語彙連鎖には、意味のない連鎖も含まれる。語彙連鎖を用いてさらに上位の文脈解析を行なうためには、生成した連鎖から意味のある連鎖 (有意な連鎖) だけを取り出さなければならない。意味のある連鎖とは、その連鎖がテキスト中で、意味のまとまりを示すものであったり、テキストの構造を示すものである。

次の条件を用いて、機械的に連鎖を取り出した。

- 章や、節のタイトルに用いられた語を含む連鎖
- 副助詞「は」「も」の直前に現れる名詞を含む連鎖
- 格助詞の直前に現れる名詞を含む連鎖
  - すべての格助詞の直前
  - 必須格を表すとされる「が」「を」「に」の直前

テキストで章・節などタイトルを持つものは、そのタイトルにしたがって記述されると考えられ、タイトルの情報を用いて主要な情報を抽出する研究が行なわれている [2]。この研究の知見から、タイトル語を含む連鎖は有意な連鎖であると考えられる。「は」など副助詞でうける語はいわいる新情報・旧情報の旧情報で、その時の話題を示すものである。このため副助詞の直前の語を含む連鎖は重要だと考えられる。格助詞がつく語は、その後の文で話題にのぼる可能性が高いので、取り上げた。

これらの条件によって取り出された連鎖が有意なものであるかどうか確認するために、まず筆者らが有意な連鎖であろうと思われるものを選別し、その連鎖と比較した。人手による選択ではできるだけ多くの連鎖を残すため、明らかに意味のない連鎖のみを削除するにとどめた。これらの条件は名詞に関するものであるので、名詞が含まれる連鎖のみを考慮した。

前出の「海の科学」第一章に対する結果では、有意な連鎖と思われるものは 69 中 46(67%)、この 46 のうち名詞の含まれる連鎖は 34 である。それぞれの条件によって取り出した連鎖の数を表 1に示す。また、タイトル語と副助詞・必須格となる格助詞の直前の名詞の含まれる連鎖のテキスト上での分布を図 2に示す。図の横一列は一つの連鎖で、左端の語は、その連鎖の分類語彙表におけるカテゴリのカテゴリ名である。縦がテキスト中の一文に相当する。文中の 1 語が連鎖に含まれる場合に#、複数の語が含まれる場合には%が付けられている。上から 4 段の連鎖がタイトル語を含む連鎖、13 段までが副助詞の直前の名詞を含む連鎖である。

条件	総連鎖数	有意な連鎖数	不要な連鎖数	欠落した連鎖数
タイトル語	4	4	0	30
副助詞の直前	13	13	0	21
格助詞の直前	33	22	11	12
必須格となる格助詞の直前	26	19	7	15
タイトル語と格助詞の直前	34	22	12	12
タイトル語と副・格助詞の直前	34	22	12	12
タイトル語と副助詞・必須格の直前	30	21	9	13

表 1: 条件により抜き出した連鎖数

5.1 考察

それぞれの条件により抜き出した連鎖には、できるだけ多く有意な連鎖を含み、不要の連鎖はできるだけ少ない方が望ましい。「タイトル語」、「副助詞の直前」は不要な連鎖を含まないものの、有意な連鎖の欠落が多い。「必須格の直前」で取り出したものが、正当率としてはもっとも高い(73%)。

しかし、それぞれの条件によって抜き出した連鎖を眺めることで、有意な連鎖の中での重要度の差を見つけることができる。タイトル語で抜き出した連鎖はテキスト全域に渡っており、このテキストの一貫性を保証する連鎖であるといえる。図2を見ると「天体」を表す連鎖がカバーする範囲は第1文から第19文までと、第31文から最後までの二つに大きく分かれている。また「宇宙・空」を表す連鎖も同じような分布をしている。このことから第20文あたりから第30文あたりまで別の話題が入り、また元の話題に戻っているという推測ができる。

タイトル語で抜き出した連鎖はすべて、「副助詞の直前」で抜き出した連鎖に包含されている。

タイトル語、副助詞の直前を含まず、格助詞の直前を含む連鎖は、テキスト中で、狭い範囲に集まっているものと広範囲に分散しているものがある。格助詞の直前を含む連鎖で不要な連鎖であるものは、主に広範囲に渡っているものであった。

6 おわりに

本研究では、語義曖昧性を解消しつつ語彙連鎖を生成するために、疑似スタック構造を用いたアルゴリズムを提案し、また、作成した連鎖から有意な連鎖を取り出すために、タイトル語情報などの情報を用いた。

今後の課題として、次のことがあげられる。結束性には語彙結束性以外にも、手がかり語に関するものもある[5]。手がかり語による話題の展開の知見を疑似スタックの制御に応用する。また、スタック中の連鎖の動きと話題の展開との関わりを調べ、照応・省略の解析に利用する方向も検討する。有意な語彙連鎖の抽出には、連鎖の濃度(連鎖がカバーしている文の数と連鎖に含まれる語の数の割合)などの定量的な手法を検討する。キーワード抽出に関する研究[7]などを参考にして、より精度の良い抽出をする方法を検討する。

参考文献

[1] Jame Morris and Graeme Hirst. Lexical Cohesion Computed by Thesaurual Relations as an Indicator of the Structure of Text. *Computational Linguistics*, Vol. 17, pp. 21–48, 1991.

[2] 高松忍, 西田富士夫. 見出し情報を用いたテキスト解析と情報抽出. 情報処理学会論文誌, Vol. 29, No. 8, pp. 760–769, 1988.

[3] 国立国語研究所. 分類語彙表. 秀英出版, 1964.

[4] 佐々木一朗, 増山繁, 内藤昭三. 結束チャートの自動生成と日本語文章の語彙的結束構造への応 用. 情報処理学会研究会資料 NL95-8, pp. 57–64, 1993.

[5] 山本和英, 増山繁, 内藤昭三. 手がかり語および語の類縁性を併用した段落分け. 情報処理学会研究会資料 NL92-6, pp. 41–48, 1992.

[6] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. 日本語形態素解析システム JUMAN 使用説明書 version 1.0, 1992.

[7] 木本晴夫. 日本語新聞記事からのキーワード自動抽出と重要度評価. 電子情報通信学会論文誌, Vol. J74-D-I, No. 8, pp. 556–566, 1991.

[8] 柳哲雄. 海の科学—海洋学入門. 恒星社厚生閣, 1988.

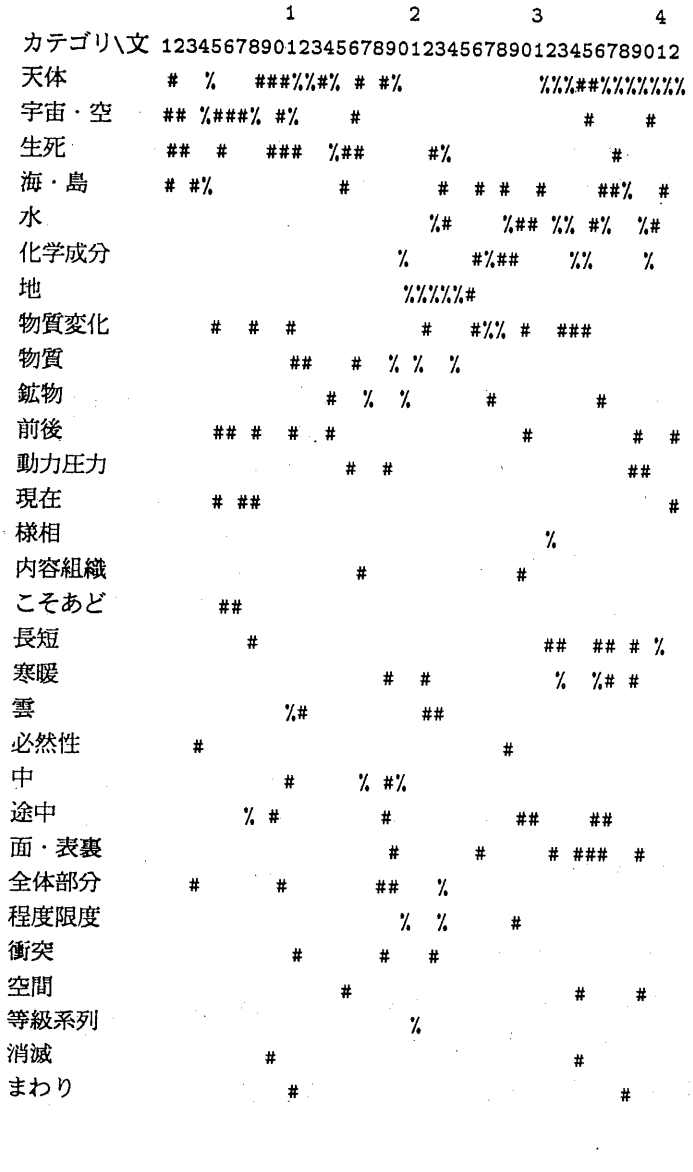

図 2: タイトル語と副助詞・必須格となる格助詞の直前の語を含む連鎖