自然言語処理 69－5
（１９８８．１２．６）

単語共起頻度を利用した形態素解析

高橋　直人　　板橋　秀一

筑波大学

本報告では、単語間の共起関係を最尤候補決定に利用する日本語文解析システムについて述べる。システムは、既に解析された部分に含まれる単語の共起情報をもとにして次に来る可能性の高い単語の範囲を予測し、その予測に従って解析を進める。これによって探索空間が狭められるので、解析速度の向上が期待できる。また、共起頻度の高い単語同士は、意味的にも強いつながりがあると考えられるので、もっともらしい解釈ほど先に得られることになる。単語間の共起頻度情報は、例文からボトムアップに学習されるので、解析の対象となる分野に合わせてシステムをチューンアップすることが可能である。

Japanese Sentence Analysis based on Word Co-occurrence Relationship

Naoto TAKAHASHI and Shuichi ITAHASHI

University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305 Japan

This paper describes a Japanese sentence analyzing system which utilizes word co-occurrnce relationship. Each time a new word is read, the system rearranges the order of retrieving words so that the word which appears most frequently with the preceeding word will be found first. Since words which are closely related one another may frequently appear in a sentence together, this strategy is expected to find the most likely result first. The system also can learn word co-occurrence relationship, so it can be tuned up by the users.

１．はじめに

計算機を用いて日本語文、特に平仮名べた書きの文を解析した場合、一般に複数の解析木が得られる。我々はこの複数の解析木の中から最尤候補を選ぶ規準として、単語間の共起頻度を用いてきた[1][2][3]。

従来のシステムでは、まずすべての可能な解析木を生成し、その中から最尤候補を選出していた。これは一定規準のもとで最高得点を持つものを選ぶ方法としては、安全かつ確実であるが、時には１０００を超える解析木をすべて生成・記憶する必要があるために、かなりの計算時間を必要とすることがあった。これら多数の解析木の大部分は、人間には思い付かないようなもので占められており、何らかの方法でそういった「非常識な」解析木を生成しない（生成した後で取り除くのではなく）ようにすれば計算コストの減少、解析精度の向上が期待できる。

人間が文を読む場合は、意味情報・文脈情報等をもとにして検索空間を限定し、それによって効率良い解析を行なっているものと考えられる。検索空間を限定するということは、次に来る単語の範囲を予想することである。LINGOL[4] は、次に来る単語の文法的カテゴリーを予測することで解析効率の向上に成功した。この方法をさらに進めて、次に来る単語の意味的カテゴリーまで予測するようにすれば、人間が文を読み進む場合に常識的な解釈を優先するのと同様の効果があるものと期待される。

今回作成したシステムは以上で述べたような方法を用いて、意味的に正しいことが期待される解析木を優先的に生成するようになっている。解析は深さ優先で進められるが、１単語読み込むごとに辞書検索の順序が変えられ、次に来る単語としては、今読み込んだ単語との共起頻度が高いものほど先に見つかるようにしている。共起頻度が高い単語同士は意味的にも近いと考えられるので、もっともらしい解析木ほど先に見つかり、「非常識な」解析木は後回しにされることになる。

２．共起頻度と形態素解析

システムの設計に当たっては、処理対象は比較的狭いものに限定し、その分野の中で意味的に妥当な文ほど速く解析できることを目標とした。逆に言えば、現在システムが持っている知識と矛盾するような文、あるいはシステムにとって「見慣れない」文の解析には、そうでないような文の解析よりも時間がかかってもかまわない、ということである。また、「意味的に近い単語同士は、共起頻度も高い」ことを仮定し、意味的整合性を共起頻度で近似した。

対象分野を限定したのは、幅広い知識を収集することが困難である、という現実的な問題のためである。後で述べるように、本システムは単語間の共起関係を例文から学習するようになっているため、知識の量・質は用意された訓練用例文に依存する。勿論、各種の類語辞典、シソーラス等を利用して、分野に依存しない、一般的な単語間の共起情報を取り込めば、比較的広い範囲の文を精度良く解析することもできると思われる。

以下では共起頻度を形態素解析にどのように利用しているかについて解説する。

形態素解析は大まかにいって

１）解析すべき文字列の先頭部分とマッチする単語を辞書の中から探す。

２）マッチする単語が見つかった場合は、その部分の文字列を、解析すべき文字列の先頭から取り除く。

３）解析すべき文字列の長さが０になるまで以上を繰り返す。

というプロセスである。本システムは、このうち１）の部分を

１－１）まず、それまでに解析された部分に含まれている単語と共起頻度の高いものの中から、未解析文字列の先頭部分とマッチするものを探す。

１－２）もし１－１）でマッチする単語が見つからなかった場合は、グローバルな辞書からマッチするものを探す。

と２段階に分けている。これによって既に解析された部分に含まれる単語と共起頻度の高い単語ほど先に見つかるようにすることができる。また、ある１つの単語と共起する単語の集合は、グローバルな辞書に含まれる単語の集合に比べてかなり小さいことが予想されるので、１－１）の段階で単語が見つかれば、初めから１－２）を行なうよりも処理が速くなる。反対に、１－１）の段階で単語が見つからなければ、余計な時間がかかる訳であるが、この場合は共起頻度の低い単語同士が結び付いていることになるので、前述の通りこれはやむを得ないものとする。

次にシステムの具体的な動作を示す。各自立語には、他の単語との共起情報がfig.1のような形で付加されている。

```
( char-1 (w-11 . f-11) (w-12 . f-12) ... )
( char-2 (w-21 . f-21) (w-22 . f-22) ... )
 ...
```

fig.1 共起情報（個別辞書）の記憶形式

char-nは１文字のキャラクタ、w-nmはchar-nで始まる単語（に与えられたid）、f-nmは、w-nmとこの情報を保持している単語との共起頻度である。現在共起頻度としては、訓練に用いた例文中で修飾・被修飾関係を伴って現れた回数をそのまま用いている。共起回数が０の単語は共起情報中に記録されていない。共起情報を持つ単語は自立語のみであるが、自立語の共起情報中には付属語も記入される。これは、「この名詞は目的語の位置に来ることが多いが、主語の位置には立ちにくい」または「この動詞は受動態で用いられることが多い」といった情報を記録できるようにするためである。なお、これ以降は「単語Ａが持っているの他の単語との共起情報」のことを、「単語Ａの個別辞書」と呼ぶことにする。

個別辞書の中では任意のnに対して

$$f\text{-}n1 \geqq f\text{-}n2 \geqq f\text{-}n3 \geqq \ldots$$

および

$$char\text{-}1 < char\text{-}2 < char\text{-}3 < \ldots$$

が成立するようにソーティングがなされている。単語を最初の１文字ごとに分類して記憶しているのは検索効率を上げるためである。（後述）

今、入力文を abcdefghijklmn という文字列とし、

word-1 = abc

word-2 = de

word-3 = fgh

という所まで解析が進んだとしよう。 次はiで始まる単語を決定するわけである。このとき、まず最初にword-3の個別辞書中の、i で始まる単語を保持しているエントリーが参照され、ijk...で始まる単語が左から順に探される。エントリー中の各単語は共起頻度順にソートされているので、共起頻度の高い単語ほど先に見つかる。

各エントリー中のサーチはリニアに行なわれるため、検索時間が心配されるが、これは実際にはほとんど問題にならない。なぜなら個別辞書中の単語が最初の１文字ごとに分類されており、そのため各エントリーに含まれる単語の数がかなり少なく押えられるからである。また、エントリー同士はそれが含む単語の先頭文字によってソートされているので、求めるエントリーの参照を高速に行なうことができる。（現在はバイナリ・サーチを用いている。）

word-3の個別辞書中にijk...という単語が見つからなかった場合は、word-2の個別辞書が調べられ、それでも見つらない場合はword-1の個別辞書が調べられる。このように、位置的に近い単語の個別辞書ほど先に調べるのは、係り受けの非交差条件を考慮したためである（次節参照）。

word-1まで遡っても見つからない時にはグローバルな辞書が参照され、それでも見つからなければその位置に未知語があるものとして解析が進められる。

```
"か" (からだ . 2) (から . 1)
     (かえる . 1)
"が" (が . 2)
"く" (くらい . 2)
"こ" (こと . 1) (この . 1)
"さ" (さえずる . 1)
"す" (すむ . 16) (する . 3)
"で" (で . 1)
"と" (と . 1)
"な" (な . 1)
"に" (に . 2)
"の" (の . 57)
"は" (は . 1)
"ま" (まねる . 1)
"や" (や . 1)
"を" (を . 1)
"オ" (オリーブ色 . 1)
"カ" (カナリア諸島 . 1)
"サ" (サギ科 . 1)
"ツ" (ツル . 1)
"移" (移る . 1) (移動する . 1)
"胃" (胃 . 1)
"一" (一種 . 1)
"羽" (羽毛 . 1)
"益" (益鳥 . 1)
"越" (越す . 1)
"黄" (黄緑色 . 1)
"回" (回る . 1)
"灰" (灰色 . 1) (灰白色 . 1)
"殺" (殺 . 1)
"活" (活動する . 1)

"観" (観賞する . 1)
"求" (求める . 1)
"去空" (去る . 2)
"空" (空中 . 1)
"形" (形 . 1)
"原" (原産 . 2)
"誤" (誤る . 1)
"口" (口 . 1)
"広" (広い . 1)
"硬" (硬い . 1)
"黒" (黒い . 2)
"最" (最大 . 2)
"子" (子 . 1)
"飼" (飼う . 1)
"似" (似る . 16)
"出" (出る . 1) (出来る . 1)
"小" (小さい . 5) (小形 . 3)
     (小型 . 2) (小鳥 . 1)
"食" (食べる . 1)
"世" (世界 . 1)
"生" (生える . 1)
"赤" (赤い . 1)
"全" (全体 . 1)
"総" (総称 . 2)
"走" (走る . 1)
"多" (多い . 1)
"代" (代表的 . 1)
"大" (大形 . 8) (大きい . 6)
     (大型 . 6) (大きな . 3)
     (大きさ . 3) (大部分 . 1)
"短" (短い . 1)

"茶" (茶色 . 1)
"中" (中形 . 6) (中型 . 6)
     (中 . 1)
"虫" (虫 . 1)
"長" (長い . 2)
"低" (低い . 1)
"渡" (渡来する . 2) (渡る . 1)
     (渡り鳥 . 1)
"動" (動物 . 1)
"特" (特産 . 2)
"禿" (禿げる . 1)
"日" (日本 . 2)
"熱" (熱帯 . 1) (熱帯産 . 1)
"白" (白い . 1) (白色 . 1)
"繁" (繁殖する . 2)
"飛" (飛ぶ . 4)
"美" (美しい . 4)
"分" (分布する . 1)
"聞" (聞かせる . 1)
"別" (別 . 1)
"変" (変える . 2)
"捕" (捕食する . 2)
"歩" (歩く . 1)
"包" (包む . 1)
"名" (名 . 2)
"鳴" (鳴く . 1)
"夜" (夜行性 . 1)
"野" (野鳥 . 1)
"来" (来る . 5)
"利" (利口 . 1)
```

fig.2 個別辞書の具体例

３．係り受け関係の決定

係り受け関係の決定は、

１）形態素解析の進行中
２）形態素解析の終了後

のいずれの場合にも試みられる。以下でそれぞれの場合について解説する。

３．１ 形態素解析進行中の係り受け解析

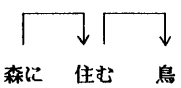
ある自立語の個別辞書によって次の単語が見つかった場合は、その２単語（厳密にはその２単語を含む文節）の間に係り受け関係のあることが期待される。このときのシステムの動作を例を挙げて説明する。

入力文を、「ブルターニュの森に住む鳥には．．．」とし、「森」および「住む」の個別辞書をそれぞれfig.3およびfig.4とする。今、文の解析が「住む」の後まで進んだとしよう。このときのシステムの状態は、大体次のようになっている。

解析済みの部分：
１．表記＝”ブルターニュの”
　カテゴリー＝名詞句
　修飾先＝名詞句
　個別辞書＝（省略）
２．表記＝”森に”
　カテゴリー＝名詞句
　修飾先＝動詞句
　個別辞書＝fig.3の内容
３．表記＝”住む”
　カテゴリー＝動詞句
　修飾先＝名詞句
　個別辞書＝fig.4の内容

未解析部分：”鳥には．．．”

前節で述べたようにまず「住む」の個別辞書中から、「鳥には．．．」とマッチする単語が検索される。今の場合これは失敗し、次に「森」の個別辞書が調べられる。ここで名詞「鳥」が見つかるので、システムは「森に」から「鳥」までの間の係り受け関係の決定を試み、各部分の文法的機能から



という係り受け関係を得る。「森に」および「住む」は係り先が決ったので、これ以上他の文節を修飾することはない（修飾を受けることはありうる）。システムは状態を

解析済みの部分：
１．表記＝”ブルターニュの”
　カテゴリー＝名詞句
　修飾先＝名詞句
　個別辞書＝（省略）
２．表記＝”森に”
　カテゴリー＝名詞句
　修飾先＝なし
　個別辞書＝fig.3の内容
３．表記＝”住む”
　カテゴリー＝動詞句
　修飾先＝なし
　個別辞書＝fig.4の内容
４．表記＝”鳥”
　カテゴリー＝名詞
　修飾先＝未決定
　個別辞書＝（省略）

未解析部分：”には．．．”

の様に変化させ、更に形態素解析を続行する。

”大”（大きい．1）
”木”（木．2）
”林”（林．2）
”高”（高い．1）
”深”（深い．3）
”繁”（繁る．3）
”鳥”（鳥．2）
”山”（山．2）（山林．2）
”神”（神社．1）

fig.3「森」の個別辞書の例

”人”（人．3）
”場”（場所．1）
”暮”（暮らす．2）
”動”（動物．1）
”棲”（棲む．2）
”都”（都．1）

fig.4「住む」の個別辞書の例

係り受け解析の結果によって、システムの状態がもっと大きく変化する場合もある。入力文を、

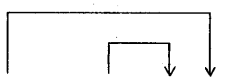
「庭でヒヨドリが柿をついばむ様子が見える」

とし、システムの状態を

1．表記＝"庭で"
　カテゴリー＝名詞句
　修飾先＝動詞句
2．表記＝"ヒヨドリが"
　カテゴリー＝名詞句
　修飾先＝動詞句
3．表記＝"柿を"
　カテゴリー＝名詞句
　修飾先＝動詞句

未解析部分："ついばむ．．．"

であるとする。また、「ついばむ」が「柿」の個別辞書中にはなく、「ヒヨドリ」の個別辞書中にはあったとする。このとき「ヒヨドリが柿をついばむ」の係り受け関係は、



の様に解析される。この場合、係り受けの非交差条件によって、「ヒヨドリ」よりも前の部分が「柿を」に係ることは決してない。そのため、文節「柿を」が他の文節によって修飾されないようにガードする必要が生じる。このときシステムは内部状態を

1．表記＝"庭で"
　カテゴリー＝名詞句
　修飾先＝動詞句
2．表記＝"ヒヨドリが"
　カテゴリー＝名詞句
　修飾先＝動詞句
3．表記＝"ついばむ"
　カテゴリー＝動詞句
　修飾先＝名詞句
　被修飾句＝［表記＝"柿を"／カテゴリー＝名詞句］

未解析部分："様子が．．．"

の様に変えて、「柿を」が外から直接は見えない状態にする。これによって非交差条件が守られる。

形態素解析進行中の係り受け解析は、文全体の情報が得られないままに起動されるため、失敗したとしてもその部分の形態素解析が間違っていることの証拠にはならない。従って、形態素解析進行中に係り受け解析が失敗した場合には、バックトラックは起こらず、単に現在の状態を保持したままで次の形態素解析に移るようにしている。

3．2 形態素解析終了後の係り受け解析

未解析の文字列が長さ0になったとき、システムは入力文全体に渡っての係り受け解析を試みる。このときの手順は、大体次のようになる。

1）係り先の決定されていない文節を、文の先頭から順にスキャンする。すべての文節の係り先が決定されている時は、係り受け解析が成功したものと見なされ、解析は終了する。そうでない時は、見つかった文節をAとする。

2）Aよりも後方の文節の中からAを受ける（Aによって修飾される）文節を探す。見つかった文節をBとする。そのような文節が見つからない時は係り受け解析は失敗である。

3）AとBに挟まれた部分

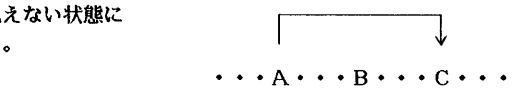
A・・・B

の係り受け関係、およびB以降の文節間

B・・・

の係り受け関係をそれぞれ再帰的に解析する。両方の解析が成功した時のみ、係り受け解析が成功したものとする。文を2つの部分に分けてそれぞれ別に解析するのは、非交差条件を満たすためである。

4）3）で行なった2種類の係り受け解析の少なくとも片方が失敗した時は、「AがBに係る」という仮定が誤っていたと考えられる。この時はBより後方の文節の中から、Aを受ける文節Cを探す。



このCを新たにBとして2）以降のステップを繰り返す。

入力文が最後まで読み込まれた後ですべての可能な組合せに対して係り受け解析が失敗したときは、形態素解析に誤りがあるものと考えられる。この場合は形態素解析をやり直し、その後で再び係り受け解析を試みるようになっている。

4．共起頻度の学習

本システムには「学習モード」があり、解析した文から単語間の共起関係情報を更新することが可能である。学習モードでは得られた解析木をユーザに提示して、それが正しいかどうかを質問する。解析結果が正しいとされれば共起情報を更新する。更新は

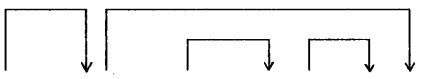
文節内：

各文節中の自立語の個別辞書を、同じ文節に現れた付属語の状況に応じて更新する。

文節間：

文節Aが文節Bに係るとき、A中の自立語とB中の自立語の共起関係をともに更新する。

のように行なわれる。例として、「オオルリの雄は鮮やかな瑠璃色で美しい。」という文が、



オオルリの　雄は　鮮やかな　瑠璃色で　美しい

の様に解析された場合の各個別辞書の更新の様子をtable.1にまとめておく。

個別辞書	頻度の更新が行なわれる単語
オオルリ	の 雄
雄	は オオルリ 美し
鮮やか	な 瑠璃色
瑠璃色	で 鮮やか 美し
美し	い 雄 瑠璃色

table.1 個別辞書とその中で共起頻度の更新が行なわれる単語

一番上の行は、「オオルリ」の個別辞書中で「の」と「雄」の共起頻度が更新されることを示す。

何をもって「共起」の規準とするかは難しい問題である。上の例では、「オオルリ」と「瑠璃色」との共起関係は更新されていないが、両者に強い結び付きがあることは明らかである。（「オオルリ」は「大瑠璃」とも書かれる。）

以前のシステムでは、1文中に現れた自立語間にはすべて共起関係があるものとしていた。この以前の方法によると、相互の依存関係がないか、あっても低い単語同士まで、共起頻度が更新されてしまうので好ましくない。

最適な規準は過去の規準と現在の規準の間にあると思われるが、これに関しては更に検討を要する。

5．インプリメントおよび実験

以上をもとにしてプログラムを書き、若干の実験を行なった。計算機はFACOM-α、言語はUtilispである。

辞書項目には、自立語類（名詞、動詞、形容詞その他）は文献5から重要語・最重要語を中心に約1万語を、付属語類（助動詞・助詞その他）は、文献6から約200語を登録している。

文法は、GPSG[7]・JPSG[8]流に各単語を素性値の集合で表す方法で表現されている。活用語尾－助動詞、助動詞－助動詞といった接続部分に見られる文法的制限も、先行単語と後続単語の素性間でマッチングをとることで実現している。

入力文には、文献5の語義解説文の中から、百科事典的説明がなされているもの（今回は特に鳥類に関する記述）を選んで用いた。

Fig.5は単語間の共起関係情報がまったく無い状態で、

「おおがたのみずべにすむとり」

という名詞句を解析した結果であり、一方 fig.6は同じ句を、例文を 100文与えて共起関係を学習させた後で解析した結果である。共起関係情報を持たない場合は連体修飾句「大形の」がその直後の名詞句「水辺」に係る、という解析結果が出ているが、共起関係学習後は「大形」の個別辞書中に「鳥」が登録されているため、正しい係り受け解析が行なわれているのがわかる。

また fig.5では「すむ」を「澄む」とする結果が、「住む」とする結果よりも先に出てきてしまっているが、共起頻度学習後は、

「水辺」と「住む」の共起頻度
＞　「水辺」と「澄む」の共起頻度

であることから、「住む」が先に選ばれている。

6．まとめ

単語共起頻度を形態素解析・構文解析に応用するシステムを作成し、若干の実験を行なった。解析した文の数がまだ少ないために定量的な評価はできないが、出現頻度の高い単語を多く含む文は、ほぼ正しく解析できると

の感触を得た。

本方式は単語対単語の共起頻度情報を利用しているので、意味素性を用いる場合よりも精度の高い解析結果が得られるものと期待される。

なお、問題点としては以下のようなものが挙げられよう。

1）常に前の単語から後の単語を予想し、その逆は行わない。そのため

例をあげる
温度をあげる

をそれぞれ

例を挙げる
温度を上げる

と正しく解析することは可能であるが、

はんきを翻す
はんきを掲げる

をそれぞれ

反旗を翻す
半旗を掲げる

と正しく解析することは保障できない。

2）形態素解析終了後の係り受け解析は、文法情報のみから決定しており、共起情報が活用されていない。

この部分は文全体の依存関係を決定するための重要な部分であり、またそれ以降の文を解析するための共起頻度にも影響を与えるので、慎重な取扱が要求される。

現在は、入力文に比較的短くて簡単なものが多いためさほど問題にはなっていないが、複雑な文を解析する場合には、共起（意味）情報・結合価情報等を利用して、より効率の良い解析を行なう必要が生じよう。

3）文より大きな単位がほとんど考慮されていない。

第2節の最後で述べたように、個別辞書のみでは引き続く部分の文字列の解析が出来ない場合には、グローバルな辞書が参照される。この場合には、最近選ばれた単語ほど優先されるようになっているが、これ以外の部分では文より大きな単位が考慮されていない。

最近の研究から、自然言語の解析のためには文脈情報の利用が不可欠であることがわかってきている[9]。本手法にも文脈情報からトピック等を抽出する機能を追加することによって、より精密な解析が可能になると思われる。

今後は以上の問題点を解決するとともに、更に大量の文の解析を行なって、システムの定量的な評価を行なう予定である。

参考文献

[1] 三浦、高橋、板橋：結合価文法における名詞のDistributionと意味素性、情報処理学会第36回全国大会4U-1（1988）
[2] 三浦、板橋、西野：結合価フレームを利用した文解析システム、情報処理学会研究報告 87-NL-63（1987）
[3] 三浦、板橋、西野：素性問題と結合価フレーム、62年度人工知能学会全国大会論文集 pp.369-37（1987）
[4] Pratt：LINGOL：A Progress Report, Proc. 4th IJCAI, pp.422-428（1975）
[5] 見坊、金田一春彦、柴田、山田、金田一京助編：新明解国語辞典（第二版）、三省堂（1974）
[6] 西村、水谷、尾上、田中：日本語基本文法単文編、電子技術総合研究所研究報告第783号、（1978）
[7] Gazdar, Klein, Pullum, Sag：Generalized Phrase Structure Grammar, Basil Blackwell（1985）
[8] Gunji：Japanese Phrase Structure Grammar, D.Reidel（1987）
[9] 堀、石崎：文脈理解とＡＩ、人工知能学会誌 Vol.3, No.3, pp.312-318（1988）
[10] 寺下、二口：分かち・構文・意味の平行処理を行なう日本語パーサ、情報処理学会研究報告 88-NL-65（1988）
[11] 尾関：最適文節列を選択するための多段決定アルゴリズム、電子通信学会技術研究報告 SP86-32（1986）
[12] 田中、辻井共編：自然言語理解、オーム社（1988）

```
% OOGATANOMIZUBENISUMUTORI
"おおがたのみずべにすむとり"

 |
 +--鳥 NIL
      |
      +--澄む (V5TAIL-U)
           |
           +--水辺に (KAKU-NI)
                |
                +--大形の (JUNTAI-NO)
AM I RIGHT (Y OR N) ? N

 |
 +--鳥 NIL
      |
      +--住む (V5TAIL-U)
           |
           +--水辺に (KAKU-NI)
                |
                +--大形の (JUNTAI-NO)
AM I RIGHT (Y OR N) ? N
```

fig.5 単語共起情報無しでの解析例

```
% OOGATANOMIZUBENISUMUTORI
"おおがたのみずべにすむとり"

 |
 +--鳥 NIL
      |
      +--住む (V5TAIL-U)
      |    |
      |    +--水辺に (KAKU-NI)
      |
      +--大形の (JUNTAI-NO)
AM I RIGHT (Y OR N) ? Y

%
```

fig.6 単語共起情報学習後の解析例