自然言語処理 72－8
（1989. 5. 19）

共起関係に注目したDM分解と確率的推定による単語のクラスタリング

松川智義
京都大学工学部
電気工学第二教室

中村順一
九州工業大学情報工学部
知能情報学科

長尾真
京都大学工学部
電気工学第二教室

要約

自然言語処理のシステムを構築する際に用いられる単語の分類を客観的なデータ解析から得る方法がいろいろと提案されている．その中には，単語の共起に関する実例データ(共起データ)を用いて単語を分類するというアプローチがある．ところが，それらの多くが前提としている単語間の「距離」(意味的な遠さ）だけで，多様な単語の意味を表現することには限界がある．また，実際の共起データには様々な「雑音」が混ざっている．

本研究では，共起データに基づいた,「距離」という考え方を用いない,「雑音」に強い，単語のクラスタリング・アルゴリズムを作成した．

An algorithm of word clustering from co-occurence data using DM decomposition and statistical estimation

Tomoyoshi MATSUKAWA
Dept. of Electrical Engineering
Kyoto University

Jun-ichi NAKAMURA
Dept. of Artificial Intelligence
Kyushu Institute Of Technology

Makoto NAGAO
Dept. of Electrical Engineering
Kyoto University

Abstract

Different methods of making word classification which is used when building natural language processing system were proposed so far. Automatic word clustering from co-occurence data is one of such approaches. However "distance" (or semantic dissimilarity) of words, which was used in most of these clustering algorithms, is not sufficient to express various meenings of words. In addition, actual co-occurence data includes several kinds of "noise".

We have developed an algorithm of word clustering which is based on co-occurence data, uses no idea of "distance" and is tough against "noise". This paper describes it.

1 序論

自然言語処理のシステムを構築する際には，単語を分類し，分類ごとに記号を与え，その記号を用いてシステムの動作を記述することが一般に行なわれている．例えば，「動詞」や「名詞」などの品詞は構文的な分類の一例である．また，非常に細かい意味記述を行なうマイクロ素性と呼ばれる記号もある[1]．比較的明確な品詞の分類であっても，単語の用法を詳細に区別しようとすると分類が不明確になる．単語の分類に対する記号の割り当てを曖昧なく安定して得ることは実際には容易なことではない．

そこで，客観的なデータ解析に基づく単語の分類法がいろいろと提案されている．その中には，単語の共起に関する実例データ(共起データ)，例えば名詞と動詞の係り受け関係のデータ，を用いて単語を分類するというアプローチがある．今までにも，因子分析により頻度付きの共起データから単語の分類を行なう方法[2]，名詞と動詞の様々な格についての共起データに基づいて単語の距離を定義し，距離空間におけるクラスタリングによって単語を分類するという方法[3]，係り受けの頻度から直接距離を定義し，その距離に基づいてクラスタリングを行なう方法[4]などが提案されている．

これらの方法に共通していることは，単語と単語の間に「意味的な遠さ」を表現する「距離」が設定できると考えていることである．このとき，複数の意味を持つ単語の場合，事前にそれらの意味が分類されていることが前提となっている．「意味」の異なる単語の間には「距離」があるはずであるからである．しかし実際に共起データを用意する場合，単語の複数の意味が分離されないままである場合も多い．この場合，単語の意味を「距離」だけで正確に表現することはできない．

一方，安定した共起データを得ることが困難であることも指摘されている[5]．共起可能性に関する人の判断には揺れがあるからである．従って，「完全に」正しく値の埋まった共起データを取り扱うことだけを考えるのは現実的ではない．ある程度「雑音」を含むデータを前提にした枠組みが必要である．

本研究では，以上のような点を考慮して，共起データに基づいた，「距離」という考え方を用いない，「雑音」に強い，単語のクラスタリング・アルゴリズムを作成した．クラスタリングは，共起データを2部グラフとみなし，その中からある程度枝を含む部分2部グラフを探し出すという形で行なう．その際，DM分解算法およびBayesの定理を利用した．

以下では，まず**2**で単語の共起関係について一般的な考察を行ない，続いて本研究で扱う問題の定式化を行なう．次に，**3**でクラスタリングのアルゴリズムを示す．最後に，**4**で，本方式による実験の結果を述べる．

2 単語の共起関係

共起関係に基づいた単語の分類を行なうためには，自然言語における単語の共起関係を観察し，モデル化する必要がある．本章では，まず**2.1**で，単語の共起関係についての一般的な考察を行なう．次に，**2.2**で，その考察に基づいて単語の共起関係をモデル化し，単語を分類するという問題を定式化する．

2.1 単語の共起関係の分類

本節では，二つの単語の共起関係を2種類に分類する．一つは，記号を用いて記述される共起関係である．もう一つは，そのような記述が可能であるが無意味な共起関係である．**2.1.1**で前者について述べ，共起関係を記号を用いて記述することの妥当性について説明する．**2.1.2**では後者について述べる．

2.1.1 記号で記述される共起関係

一般の分類とのずれ

自然言語処理に限らず，単語の分類は様々な形で行なわれている．しかし，自然言語処理に使われる単語の分類と一般に行なわれている単語の分類との間には「ずれ」がある．例えば，一般に単語の分類として定着している「品詞」を日本語の統語処理に対する必要性の点から見ると，

- 名詞と動詞の区別は重要
- 形容詞，形容動詞，動詞の区別はあまり重要ではない

などの違いがある．また，次の例のように，「犬科」「猫科」などの動物学上の分類が統語処理には使えないということはよく知られている．

例 1 犬が吠える．
ライオンが吠える．
狐が鳴く．
猫が鳴く．

従って，自然言語処理においては，一般に行なわれている単語の分類をそのまま用いることはできない．

本論文では，自然言語処理に用いられる単語の分類のうち，単語と単語の係り受け[1]を判断する処理(広義の統語処理)に限定して議論をすすめる．

共起関係の記述

同じ振舞いを示す単語の集合を，ここでは仮にカテゴリと呼ぶことにする．二つのカテゴリがあって，その各々に属する単語の全ての組合せについてある係り受けが可能であるとき，そのカテゴリを用いてその係り受けを記述することができる．例えば，「紐」「糸」「縁」「手」「話」にカテゴリ記号≪紐≫を，「切る」「結ぶ」「繋ぐ」にカテゴリ記号≪紐操作≫を割り当てることにより，

例 2 紐/糸/縁/手/話　を　切る/結ぶ/繋ぐ.

という共起関係を

例 3 ≪紐≫　ヲ　≪紐操作≫

と記述できる．同様に，「構造」「家」「機械」「話」「計画」にカテゴリ記号≪構造≫を，「作る」「壊す」「潰す」「修復する」にカテゴリ記号≪構造変更≫を割り当てることにより，

例 4 構造/家/機械/話/計画　を　作る/壊す/潰す/修復する.

という共起関係を

例 5 ≪構造≫　ヲ　≪構造変更≫

と記述できる．

実際の共起データ

実際に共起データを集めると，人による判断の揺れや慣用の変化などによって必ずしも安定したデータが集まらない．この場合，共起関係をカテゴリの対で記述すると例外が含まれることになる．例えば，**例 2**において，

縁/話　を　繋ぐ

の組み合せが「係り得ない」と判断された場合，実際に得られるデータは

例 6 紐/糸/手　を　切る/結ぶ/繋ぐ.
縁/話　を　切る/結ぶ.

となり，**例 3**のように記述すると，「縁/話を繋ぐ」は例外となる．

[1]本研究では係り受けを，名詞と自動詞/形容詞/形容動詞の「ガ格」に関する係り受けと，名詞と他動詞の「ヲ格」に関する係り受けに限定して議論する．また，基本的には，動詞/形容詞/形容動詞には受動や使役，テンス，アスペクトなどは付加されないものとする．

意味素性による記述

上述のような共起関係の記述を行なった場合，一種類の共起関係につき二つのカテゴリ記号が与えられる．例えば，**例 3** の場合には，≪紐≫，≪紐操作≫というカテゴリ記号が，**例 5** の場合には，≪構造≫，≪構造変更≫というカテゴリ記号がそれぞれ与えられている．このようなカテゴリ記号を用いて**例 3**，**例 5** のように記述するのは，≪名詞≫，≪他動詞≫などのカテゴリ記号を用いて，

例 7　≪名詞≫　ヲ　≪他動詞≫

などと記述するのと同様，自然なことである．

しかし，記号の与え方という観点から考えると，少し違った方法も考えられる．例えば，単語の集合に記号を割り当てるのではなく，共起関係の種類ごとに記号を割り当てるという方法が考えられる．この場合，与えられる記号は一種類の共起関係につき一つですむ．例えば，**例 2** の場合，名詞「紐」「糸」「縁」「手」「話」と動詞「切る」「結ぶ」「繋ぐ」にすべて同じ記号≪紐-紐操作≫を与える．同様に，**例 4** の場合，名詞「構造」「家」「機械」「話」「計画」と動詞「作る」「壊す」「潰す」「修復する」にすべて同じ記号≪構造-構造変更≫を与える．このとき，名詞「話」のように複数個の記号が与えられることも許すとする．そして，二つの単語の共起可能性を判断するときには，

1. 同じ記号が与えられていれば，共起する
2. そうでなければ，共起しない

とすることにする．別に**例 7** のような粗い統語規則が用意されてあるとすれば，このようにしても共起関係は記述できる．以降，本研究ではこのような共起関係の種類ごとに与えられる記号を**意味素性**と呼ぶことにする．

2.1.2 特異な共起関係

前節で述べたように，共起関係にはカテゴリの対で記述できる場合がある．さらに，ひとつの単語しか含まないようなカテゴリも許すとすれば，全ての共起関係がカテゴリの対で記述できる．例えば，

例 8　身の毛　が　よだつ．

という共起関係の場合，名詞「身の毛」にカテゴリ記号≪ミノケ≫を，動詞「よだつ」にカテゴリ記号≪ヨダツ≫をそれぞれ割り当てることにより，

例 9　≪ミノケ≫　ガ　≪ヨダツ≫

と記述できる．「身の毛」も「よだつ」も現代の日本語においてはほとんどこの組合せでしか使われない単語であるが，その場合でもカテゴリの対で記述できることになる．

しかし，記述の効率の点から考えると，**例 3** などの場合に比べて**例 9**の場合は，カテゴリ記号を導入する利点があまりない．**例 3** の場合は，二つのカテゴリ記号(あるいは，一つの意味素性)で**例 2** の15通りの共起関係が記述されていた．それに対して**例 9** の場合は，**例 8** の1通りの共起関係しか記述されていないからである．このような場合，動詞「よだつ」の辞書に直接，

例 10　よだつ　(ガ：身の毛)

と記述する方が自然である．

この例などから考えると，共起関係をカテゴリの対で記述する場合は，両方のカテゴリが少なくとも二つ以上の単語を含む必要があることがわかる．**例 9** のように，カテゴリが一つしか単語を含まない場合は，個々の単語の辞書に直接共起関係を記述しても大差ないからである．以降，**例 8** のような共起関係を**特異な共起関係**と呼ぶことにする．

2.2 問題の定式化

二つの品詞の共起関係をすべての単語の組合せについて表にすると図1のようになる．このような表を本研究では**共起表**と呼ぶことにする．共起表は二つの単語集合（品詞）について，その要素のすべての組合せの共起可能性を「係り得る(1)」「係り得ない(0)」の2値で表現したものである．これは，グラフ理論的に見ると，2部グラフとみなすことができる(図2)．

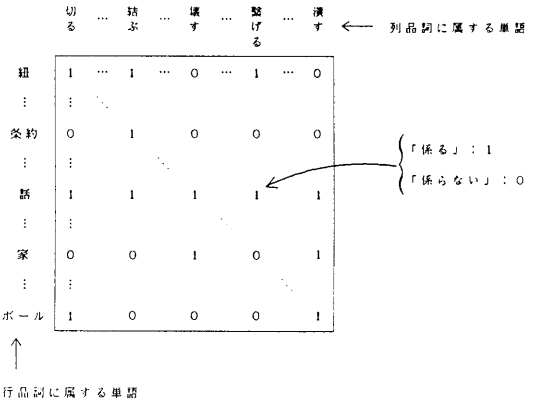


図 1: 共起表

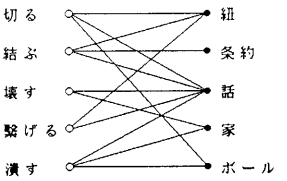


図 2: 2部グラフの例

一方，前節で述べた**意味素性**は品詞の部分集合の対であった．これは，共起表における完全部分2部グラフに相当する．したがって，意味素性を設定するという作業は，共起表の中から適当な完全部分2部グラフを探し出すという問題に置き換えることができる．以下では，グラフ理論の用語を用いて，共起データから意味素性を抽出し単語を分類するという問題を定式化する．

2.2.1 用語の定義

共起表: V^+, V^-を2部節点集合とする2部グラフ．ただし，列品詞V^+及び行品詞V^-はどちらも単語の集合であるとする．

以下では，特に断わりのない限り共起表を，

$$T = (V^+, V^-, E)$$
$$\text{ただし}, E \subseteq V^+ \times V^-,$$

と表わすことにする．ここでEは枝集合を表わす．

共起ペア: 枝集合Eに属する単語ペア．

「共起ペアである」と「単語ペアに値1が振られている」は同じことを意味する．以降，この二つの表現を場合によって使い分ける．

格子: 共起表$T = (V^+, V^-, E)$の誘導部分2部グラフ，

$$c = (U^+, U^-, D)$$
$$\text{ただし}, U^+ \subseteq V^+, U^- \subseteq V^-, D = U^+ \times U^- \cap E$$

で

$$|U^+| \geq 2, \text{かつ}, |U^-| \geq 2$$

を満たすもの．

格子の大きさ: 格子の節点集合に属する単語の個数．

すなわち，格子 $c = (U^+, U^-, D)$ の大きさ $|c|$ は，

$$|c| = |U^+| + |U^-|$$

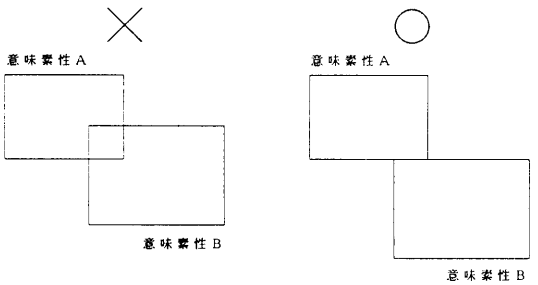
である．

2.2.2 意味素性の仮定

本節では，共起表における共起ペアの配置に関する仮定である**意味素性の仮定**について述べる．この仮定は，共起表の理想的な姿を表現したものである．この仮定を満たす共起表を，**完全共起表**と呼ぶ．**意味素性**は，完全共起表における完全格子のうちのあるものであるとする．

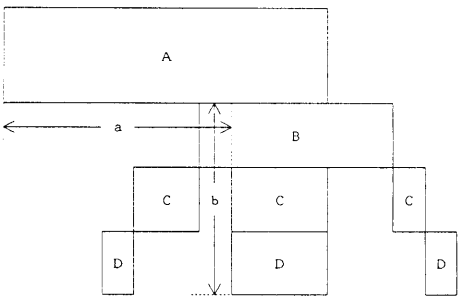
意味素性の仮定

意味素性の仮定1は，一つの単語ペアが二つ以上の意味素性に含まれることがないという制約である(図3)．意味素性が「意味解釈」の種類を表現していると考えると，この制約は二つの単語の共起の「意味解釈」が常に一つであるということを表現していることになる[2]．



意味素性は互いに単語ペアを共有することはない．

図3: **意味素性の仮定1**



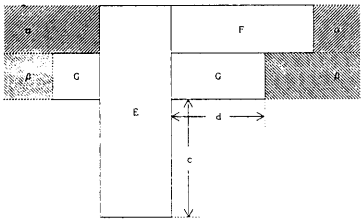
a ＞ b

図4: **意味素性の仮定2**

[2]自然言語の共起データにはこの制約を満たさない例も存在する．例えば，

例 足を洗う．

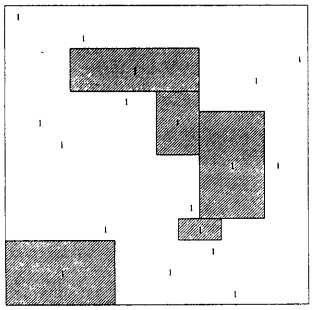
のような，直接的な意味と比喩的な意味の二つの「意味解釈」が可能な場合がそれである．しかしこの場合も，比喩的な意味の方を特異な共起とみなすとすると，意味素性については仮定が満たされているとも言える．



c ＞ d

αあるいはβに意味素性は存在しない．

図5: **意味素性の仮定3**



意味素性に含まれない1は各行各列に高々一つ

図6: **意味素性の仮定4**

次に，**意味素性の仮定2**，**意味素性の仮定3**は，意味素性の共起表における配置と相対的な大きさに関する制約である．**意味素性の仮定2**は例えば，意味素性A，B，C，D(ただし，$|A| \geq |B| \geq |C| \geq |D|$ とする)が図4のように配置しているとすると，長さ a，b について，

$$a > b$$

であるという制約である．また，**意味素性の仮定3**は例えば，意味素性E，F，G(ただし，$|E| \geq |F| \geq |G|$ とする)が図5のように配置しているとすると，長さ c，d について，

$$c > d$$

であるという制約と

α あるいは β には意味素性は存在しない．

という制約である．

最後に，**意味素性の仮定4**は，特異な共起が一つの単語につき高々一つしか存在しないという制約である(図6)．すなわち，ある単語に注目したとき，その単語を含む共起ペアは一つを除いてすべて意味素性に含まれることになる．これは，ほとんどの共起関係が意味素性で説明できるということを表現している．

これらの仮定を導入することによって，完全共起表における意味素性の解釈(つまり，どの完全格子を意味素性とみなすか)が一意に決まることが保証される(証明略)．

2.2.3 完全共起表に対する雑音

自然言語の共起データの不規則性をモデル化するために，完全共起表に対する**雑音**という考え方を導入する．

雑音

自然言語の共起データが不規則である原因には二つある．一つは**2.1.1**でも述べたように，意味素性で記述すると一部に例外が生じるような共起関係があるということである．その中には，

意味的な不整合や慣用的な制限によって生じる例外やデータを用意するときの判断の揺れによる例外などがある．もう一つは，特異な共起関係が，**意味素性の仮定4**に比べて，数が多かったり分布が偏っていたりするということである．

これらの不規則性は共起表の上では以下のように表現できる．

まず，前者の意味素性に例外が含まれる場合は，完全共起表において本来完全格子であるはずの意味素性が完全格子ではなくなっていると表現することができる．意味素性に含まれる一部の単語ペアの値が1から0に反転していると考えるのである．以降，このような単語ペアの値の反転を**負の雑音**と呼ぶことにする．またこのとき，意味素性中で単語ペアの値が1から0へ反転している割合を**1-0反転率**と呼ぶことにする．

一方，後者の特異な共起関係の増大/偏りは，逆に，本来0となるべき単語ペアの値が1に反転していると表わせる．以降，このような単語ペアの値の反転を**正の雑音**と呼ぶことにする．また，値が0である単語ペアのうち値が0から1へ反転している割合を**0-1反転率**と呼ぶことにする．

以上より，自然言語の共起データは，完全共起表に負の雑音，正の雑音がそれぞれ1-0反転率，0-1反転率で加わったものとみなすことができる．また逆に，共起データからこれらの雑音を除去すれば，完全共起表が得られると考えられる．そしてその際，単語の分類が意味素性として抽出されることになる．

本研究では，完全共起表を復元し意味素性を抽出する作業を，単語を分類する作業とみなすことにする．

3 完全共起表の復元

完全共起表を復元する手続きは意味素性を復元する手続きを繰り返すという形で実現する．意味素性を復元する手続きは次の2つの部分に分かれる．

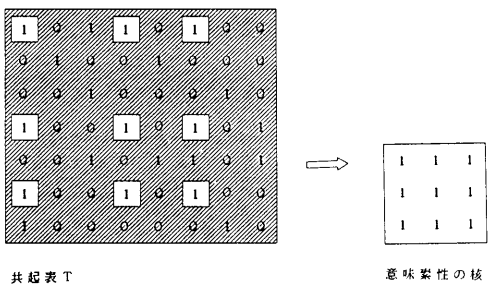


図7: 意味素性の核を求める

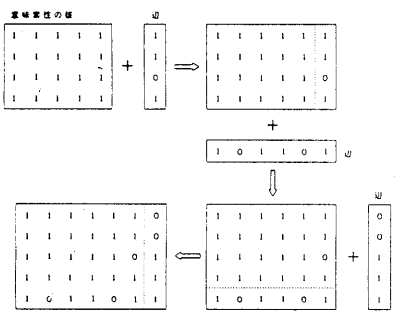


図8: 辺を付加していくことによって意味素性の核を成長させる

1. まず，共起表中で最も大きい完全格子(意味素性の核)を求める(図7)．
2. 次に，値1がある程度埋まった辺[3]を付加していくことによって，意味素性の核を成長させる(図8)．

以下では，**3.1**，**3.2**で，それぞれ1，2の部分について説明する．続いて**3.3**で，全体のアルゴリズムについて述べる．

3.1 意味素性の核を求める

2部グラフの完全部分2部グラフの節点集合は，その補2部グラフの独立節点集合である(図9)．2部グラフの独立節点集合の補集合は，そのグラフの点被覆である(図10)．従って，2部グラフの最大完全部分2部グラフ(の節点集合)は，補2部グラフの最小点被覆の補集合に相当する．2部グラフの最小点被覆を求める算法としては$O(n^2)$の計算量のDM分解算法が知られている[6][7]．従って，DM分解算法を用いることにより，2部グラフの最大完全部分2部グラフを求めることができる．

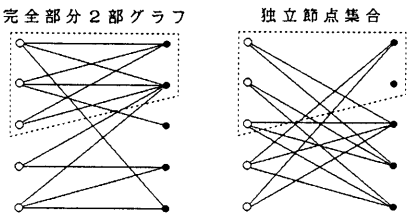


図9: 2部グラフの完全部分2部グラフの節点集合は，補2部グラフの独立節点集合である．

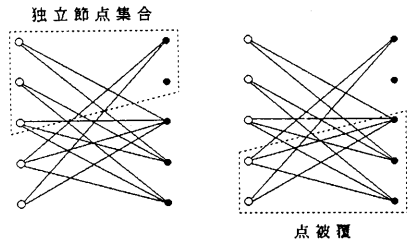


図10: 2部グラフの独立節点集合の補集合は，点被覆である．

単語ペアを含む最大の完全格子を求める手続き <find-complete-check(p,T)> は，これを利用している．以下にそれを示す．なお，この手続きの中のステップ1で求めている同類共起表T_pというのは，単語ペアpを含む完全格子をすべて含む部分共起表である[4]．また，ステップ4でKの中から1つ選んでいるが，Tが完全共起表の場合Kの要素数は必ず1つであるのでこれは一意に決まる(証明略)．一方，Tが完全共起表でない場合は，Kの要素は複数になり得るが，この場合は任意に1つ選ぶとする．

[3]共起表の誘導部分2部グラフのうち，片方の品詞の単語数が1であるものを辺と呼ぶ．

[4]単語ペア$p=(v_0^+,v_0^-)$，及び，共起表$T=(V^+,V^-,E)$について，Tの誘導部分グラフ

$$T_p=(V_p^+,V_p^-,E_p)$$
$$\text{ただし},V_p^+\subseteq V^+,V_p^-\subseteq V^-,E_p=V_p^+\times V_p^-\cap E$$

で，

$$V_p^+\times\{v_0^-\}\subseteq E_p,$$
$$\{v_0^+\}\times V_p^-\subseteq E_p,$$
$$\forall v^+\in V_p^+:\ |\{v^+\}\times V_p^-\cap E_p|\geq 2,$$
$$\forall v^-\in V_p^-:\ |V_p^+\times\{v^-\}\cap E_p|\geq 2$$

を満たすもののうち，大きさが最大のものをpのTにおける**同類共起表**と呼ぶ．

手続き:単語ペアを含む最大の完全格子を求める<find-complete-check(p,T)>

入力: 共起ペア $p=(v_0^+, v_0^-)$,
共起表 $T=(V^+, V^-, E)$
出力: 完全格子 c (ただし,c は p を含む)
保証: T が完全共起表で p が意味素性に含まれていれば，c は p を含む最大の完全格子である．
計算量: $O((|V_p^+|+|V_p^-|)^2)$

1. p の T における同類共起表 $T_p=(V_p^+, V_p^-, E_p)$ を作る．
2. T_p の補2部グラフ $\bar{T}_p=(V_p^+, V_p^-, V_p^+ \times V_p^- - E_p)$ を作る．
3. $\bar{T}_p$ のすべての最小点被覆(n個あるとする)をDM分解算法によって求め，その集合を

$$\{(\bar{W}_i^+, \bar{W}_i^-) | i=1,2,\cdots,n\}$$

とおく．このとき，T_p のすべての最大完全部分2部グラフ

$$K=\{(W_i^+, W_i^-, W_i^+ \times W_i^-) | i=1,2,\cdots,n\}$$
ただし, $W_i^+ = V_p^+ - \bar{W}_i^+, W_i^- = V_p^- - \bar{W}_i^-$

が同時に求まっている．

4. K のうち，

$$|W_i^+| \geq 2, \text{かつ}, |W_i^-| \geq 2,$$

を満たすものがあればその中から任意に1つ選び，出力として終了する．なければ，nil を出力として終了する[5]．∎

この手続きをすべての共起ペアに適用することによって，意味素性の核を求める．

手続き:意味素性の核を求める <find-sf-nucleus(T)>

<find-complete-check(p,T)>をすべての共起ペアに適用し[6]，得られた格子のうち最大のものを求めれば，共起表中の最大の完全格子(意味素性の核)が得られる．∎

3.2 意味素性の核を成長させる

互いに排他的な事象の系列 B_i(ただし，$\sum_i B_i = \Omega$) があるとき，事象 A についての B_i の条件付き確率を求める式，

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

は一般に，Bayesの定理と呼ばれている．本研究では，この定理の事象 A，B_i に

A : 雑音が加えられた共起表中の長さaの辺が，値1をb個含む．
B_i : その辺が，元の完全共起表において，値1をi個含む．

をそれぞれ割り当てて，$P(B_a|A)$ を与える式を導出した(詳細略)．この式は，共起ペアを b 個含む長さ a の辺があるときに，「それが元の完全共起表において完全格子の一部であった．」という推定の信頼度とみなせる．従って，この値が大きいほど，意味素性の核に付加する辺としての適合度が高いと考えられる．以下に，導出された適合度を与える式を示す．

手続き:格子に加えられる辺の適合度 <suitability(s,T)>

入力: 辺 $s=(W^+, W^-, B)$
ただし，$|W^+|=1$ あるいは $|W^-|=1$，
$B = W^+ \times W^- \cap E$
出力: 「$W^+ \times W^- \subseteq E_0$であった」という推定の信頼度 R
ただし，E_0 は完全共起表の枝集合

$|W^+| \times |W^-| = a$，$|B| = b$ とおく．1-0反転率，0-1反転率をそれぞれ q，r とおく．すると，「$W^+ \times W^- \subseteq E_0$であった」という推定の信頼度 $R(a;b)$ は次式で与えられる，

$$R(a;b) = \frac{{}_aC_b(1-q)^b q^{a-b}}{{}_aC_b(1-q)^b q^{a-b} + A(a,b)}$$

ただし，

$$A(a,b) = \frac{1}{2^{a-1}} \sum_{i=0}^{a-1} {}_{a-1}C_i \sum_{j=0}^{i} B(a,b,i,j)$$

$$B(a,b,i,j) = {}_iC_j(1-q)^j q^{i-j} \cdot {}_{a-i}C_{b-j} r^{b-j}(1-r)^{(a-i)-(b-j)}.$$ ∎

上記の適合度に基づいて意味素性の核に付加していく辺を選び，それを繰り返すことにより意味素性の核を成長させていく．

手続き:意味素性の核を成長させる <growing-up($nucl$,T)>

格子 $nucl$ に付加できる辺のうち **<suitability(s,T)>** が最大のものを任意に1つ選び，その値が0.5以上ならば，それを $nucl$ に付加し，新たに $nucl$ とする．
以上を，付加する辺がなくなるまで続ける．∎

3.3 全体のアルゴリズム

全体のアルゴリズムを以下に示す[7]．

手続き:完全共起表を復元する <restore-complete-table(T)>

入力 : 共起表 $T=(V^+, V^-, E)$
出力 : 共起表 $T'=(V'^+, V'^-, E')$
保証 : T が完全共起表のとき，$T'=T$
計算量 : $O(nm^2|E|)$
ただし，n: 意味素性の個数，
m: 意味素性の大きさ

```
T' := T
while nucl := <find-sf-nucleus(T)> do
    c := <growing-up(nucl, T)>
        ;nuclをT'の中で成長させる．
    (U+, U-, D) := <remove-unsuitable-words(c)>
        ;不適切な辺を取り除く．
    E := E - U+ × U-
        ;U+ × U-の要素をすべてTの枝から取り除く．
    E' := E' ∪ U+ × U-
        ;U+ × U-の要素をすべてT'の枝に加える．
od
return(T').
```
∎

[5]「単語ペアを含む最大の完全格子」を正確に求めるならば，ここは nil とするのではなく，さらに細かく分解していく必要がある．しかし「意味素性に含まれる共起ペア」については「共起ペアを含む最大の完全格子」がこの手続きで必ず求められるので(証明略)，ここは nil で十分である．

[6]分枝限定法により，一部の共起ペアへの適用を省略することは可能である．

[7]ただし，この中で使われている<remove-unsuitable-words(c)>は，**<suitability(s,T)>** を c のすべての辺に適用し，その値が0.5以下のものを c から取り除く手続きである．

表 1: 機械的に作成した共起表の場合

q	r	A(14,19)	B(18,6)	C(13,11)	D(9,7)	E(5,4)	復元率(面積)
0(%)	0(%)	○	○	○	○	○	
		100(%)	100(%)	100(%)	100(%)	100(%)	100(%)
1	0	○	(17,6)	○	○	○	
		100	94.4	100	100	100	98.8
5	0	○	○	○	○	○	
		100	100	100	100	100	100
10	0	○	○	(12,11)	○	(4,4)	
		100	100	92.3	100	60.0	97.5
20	0	○	(17,6)	(13,9)	(6,7)	(5,3)	
		100	94.4	81.6	66.8	75.5	92.6
30	0	(14,17)	(14,6)	(13,10)	(7,5)	○	
		69.4	77.7	90.9	55.5	100	84.5
40	0	(13,16)	(14,6)	(11,10)	(6,7)	(4,3)	
		76.1	77.7	76.9	66.6	60.0	76.3
50	0	(12,18)	(12,4)	(11,9)	(7,6)	(4,3)	
		81.2	44.4	69.2	66.6	60.0	69.5
60	0	3つに	2つに	(10,10)	(5,5)	(4,2)	
		分離(度)	分離(度)	60.9	39.6		20.8*
1	5	○	○	○	○	(4,4)	
		100	100	100	100	60.0	99.3
5	5	○	○	(12,10)	○	(4,2)	
		100	100	83.9	100		92.8*
10	5	○	○	2つに	○	(4,2)	
		100	100	分離(度)	100		72.6*
20	5	(13,18)	(14,6)	2つに	(9,5)	○	
		92.6	77.7	分離(度)	71.4	100	66.0*
30	5	(13,15)	(13,6)	(12,10)	(7,6)	(4,2)	
		73.3	72.2	63.9	66.6		72.5*
40	5	(12,17)	(14,6)	(13,10)	(7,6)	(4,2)	
		76.6	77.7	90.9	66.6		76.6*
1	10	○	(17,6)	○	(9,6)	(4,3)	
		100	94.4	100	85.7	60.0	96.1

*は(度)の場合を計算に入れていない

4 実験結果

<restore-complete-table(T)> を

1. 機械的に作成した共起表
2. 自然言語の共起データ

の2つの場合に適用する実験を行なった．以下の節で，これらについて報告する．

4.1 機械的に作成した共起表

実験の設定

- 完全共起表を(意味素性の仮定を満たすように)作る．
  その際意味素性は，

$$A(14\text{語} \times 19\text{語}), B(18\text{語} \times 6\text{語}), C(13\text{語} \times 11\text{語}),$$
$$D(9\text{語} \times 7\text{語}), E(5\text{語} \times 4\text{語})$$

  の5つとする．
- 共起表全体に，1-0 反転率=q(%)，0-1 反転率=r(%) でそれぞれ負の雑音，正の雑音を加える．
- <restore-complete-table(T)> のパラメータ q, rは，正しい値がわかっているとして，上記の q, rをそのまま使う．

実験結果

実験結果を表1に示す．ここでは，復元された意味素性が，例えば(45,7)などと表わされている．これは，意味素性のうち45語×7語が復元されたことを意味している．また，意味素性が完全に復元された場合は，○で表わしている．各欄の下には，意味素性の復元率が単語ペアの個数比(格子の面積比)で示されている．

この結果をまとめると次のようになる．

1. $q = 0$(%)，$r = 0$(%)の場合は，理論通りすべての意味素性が抽出できた．
2. $r = 0$(%)の場合，
   (a) $q \leq 50$(%)の範囲では，意味素性が混ざって抽出されることはなかった．
   (b) $q = 50$(%)の場合でも，意味素性の約70%(面積比)が復元できた．
3. $r = 5$(%)の場合，
   (a) $q = 10$(%)ですでに分離して抽出される意味素性が出てきた．

結果の検討

以上の結果から，次のことがわかる．

- 負の雑音の変化にはあまり影響されない．
  従って，共起データの多少の不備に対しては，本方式は十分に対処できる．
- 正の雑音の変化には敏感に影響される．
  従って，特異な共起関係が予想よりも多くなった場合，本方式の自然言語データへの適用は困難になる．

4.2 自然言語の共起データ

実験の設定

- テニスの入門書から名詞100語，動詞50語を選ぶ．
- そのすべての共起関係を筆者自身が判定して共起表を作る．
- <restore-complete-table(T)>のパラメータ q, rは，それぞれ $q = 20$(%)，$r = 0$(%)と仮定する．

実験結果

実験結果を図11に示す．この結果をまとめると次のようになる．

- 好ましい結果
  - カテゴリとみなし得る単語集合が意味素性の一方の節点集合に現れた．
  - 「うまい」「まわる」「曲がる」「力強い」「ボレー」などの意味の分離に成功した．
- 好ましくない結果
  - 比較的大きい意味素性には，いくつかのカテゴリが混ざっているようであった．
  - 特異な共起がかなり残った．
  - 「あがる」「フォア」などの意味の分離は失敗した．

結果の検討

この実験結果より，次のことが言える．

- カテゴリと考えることができる単語集合がある程度抽出できていることから，「共起表における適当な完全格子を探す」という基本的な方針は有効であると思われる．
- 人間には気がつきにくい単語の細かい意味分類を示唆するような結果が出ているので，辞書構築の際の基礎データを提供する方法として有望であると思われる．
- 抽出された意味素性のうち大きい意味素性は，いくつかのカテゴリを混在させていたので，「大きい完全格子から意味素性としていく」という方法は，検討の余地があると言える．
- 特異な共起ペアが多く残ったことから，
  - 共起データの準備方法
  - モデルと現実とのずれ

  の両面から検討を加える必要があるということが言える．

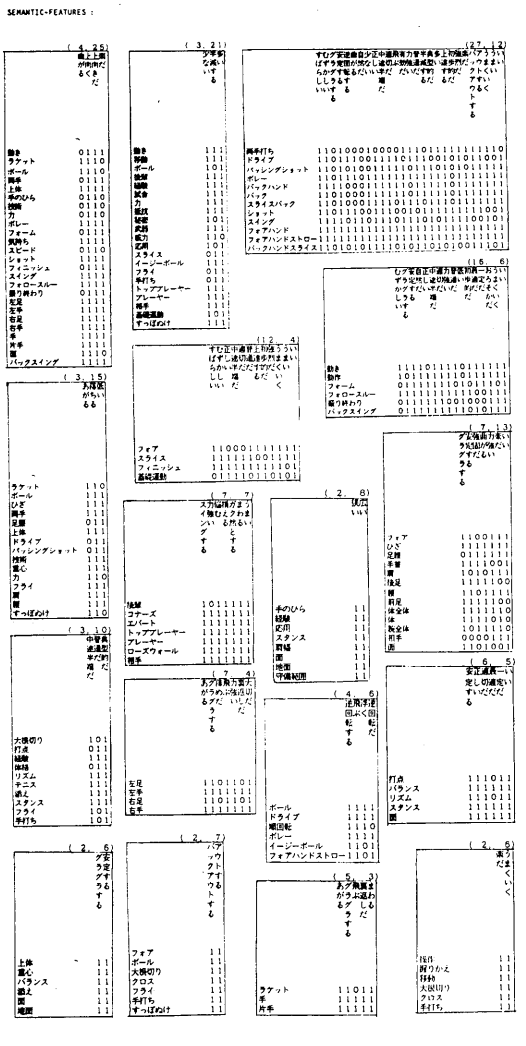


図 11: 自然言語の共起データの場合

5 結論

本研究では，単語の共起関係を辞書や文法に記述する際に用いられる「意味素性」について考察し，それを共起データから得るための方法を示した．その際，二つの品詞の共起関係の理想的な姿を「完全共起表」という形でモデル化し，実際に得られる共起データを完全共起表に「雑音」が加わったものとして表現した．そして，雑音を取り除き完全共起表を復元することで意味素性を得るという問題設定を行ない，その問題を解くアルゴリズムを作成した．

完全共起表に様々な割合の雑音が加わったものについてシミュレーション実験を行なうことで，作成したアルゴリズムの妥当性を実証した．また，自然言語のデータに対しては，名詞100語×動詞50語の「ガ格」についての共起データを用いて実験を行ない，今後の参考とした．その結果，

1. 完全共起表からは常に正しく意味素性が自動抽出できる．
2. 完全共起表中の値1が一部欠けている場合，欠けている割合が50%以下であれば，70～100%の正解率で意味素性が自動抽出(復元)できる．
3. 完全共起表に比べて特異な共起が多い場合，特異な共起の割合が5%を越えると意味素性の自動抽出が困難になる．

ということがわかった．今後の課題としては，

1. 大規模な自然言語の共起データを用いて本モデルの妥当性及び本方式の有効性を検討すること．
2. 統語解析・生成システムと結合することにより，求められた意味素性の有効性を実証すること．

が挙げられる．

参考文献

[1] 丸山　宏，渡辺日出雄．既存のメニューシステムと自然言語インターフェースの融合．日本ソフトウェア科学会第4回大会論文集，63-66，1987．

[2] 仲尾由雄，栩山陽子．係り関係による単語のクラスタリングの試行．情報処理研究報告 *NL65-1*，1988．

[3] 藤原祥隆．共起パタン分布に基づく単語間類似度を用いた動詞・名詞のクラスタリング法．第2回人工知能学会全国大会論文集，449-452，1988．

[4] 白井克彦，林良彦，平田裕一，久保田淳市．係り受け解析のための辞書の構成とその学習機能．情報処理学会論文誌 *Vol.26 No.4*，706-714，1985．

[5] 田中英輝，江原暉将．日本語「が」格の係り受け整合度の一計算法．情報処理学会研究報告 *NL70-10*，1989．

[6] 伊理正夫　他共著．演習グラフ理論．コロナ社，1983．

[7] 服部嘉雄，小澤孝夫　共著．グラフ理論解説．昭晃堂，1974．