情 報 学 基 礎 66－6
デジタル・ドキュメント 32－6
（２００２．３．１５）

複合語からの類義語抽出法

中渡瀬 秀一†

本論文ではテキストデータを解析して，そこから類義語のグループを自動抽出する方法を提案する．提案する方法ではテキストを形態素解析した情報より２部グラフを構成し，その中の完全２部グラフが類義語のグループを構成するという傾向を利用して類義語の自動抽出を行う．この方法は大量のデータから多くの類義語候補を容易に収集するのに有用である．また２部グラフを構成する双方の頂点集合が類義語グループとそのグループの観点として解釈できるため，類義語のグループと同時にその観点も構成することができる．さらにグラフ間の順序関係によって類義語間の類似度を与えることもできるという特徴をもつ．

A Method for Extraction of Similar Words from Compound Nouns based on Complete Bipartite Graph

HIDEKAZU NAKAWATASE †

This paper proposes a method for similar words groups extraction from nouns in a corpus. This method is based on maximal complete bipartite graph included bipartite graph made from compound nouns. In this bipartite graph, one node set tends to be a similar words group and the other node set to be a group of view point words, when the graph is complete. Further, this method can give also similarity measurements between words using relations among those similar words groups.

1. はじめに

本論文では２部グラフ構造を利用して，語彙間の関係の集合から自動的に類義語グループを抽出する方法を提案する．

情報検索など自然言語処理を必要とするアプリケーションにおいて、語彙間の類似度を決定したり，ある語彙に類似する語彙の集合を生成するために用いられる類義語辞書の役割は重要である．例えば情報検索では適当な検索語集合を獲得するための辞書として用いられたり，与えられた検索語の類義語そこから獲得して検索式を拡張するのに用いられる．

しかし伝統的に類義語辞書の構築は大部分が人手によって行われる非常に手間のかかる作業であった．なぜならＮ個の語彙があるとき各語彙について複数の観点でグループ化することも考えると膨大な組み合わせの判定を行わなければならないからである．実際には分類体系の大きさ (グループの数) と分類体系の構造をあらかじめ方針として定め試行錯誤的に著者が設計した体系を用いて語彙をグループ化している[2]．

単語辞書に収録する語彙は文献から抽出することによって得られるので，収集元となる文献を増大させることによって労力のある限り辞書を充実させることができる．ところが分類体系の方は創作物であり分類項目 (語彙のグループ) をどう選択し構造化してゆくかには任意性があるため明確な基準を定めることが難しい．伝統的な類義語辞書[1] などでは木構造，上位下位関係が使用されてきた．しかし言語処理にこのような類義語辞書を使用する際の問題点は川村[4] が指摘しているように複数の観点によって語彙間の関係を表現するのが難しい点にある．

そこで筆者は多観点の類義語辞書を作成するために必要な類義語グループ候補を自動生成する方法、その類義語間に類似度を与える方法を提案する．提案する方法では予めテキストを形態素解析して得られた語彙間の関係より２部グラフを構成し，その中の完全２部グラフが類義語のグループを構成するという傾向を利用して類義語グループの自動抽出を行う．この方法は大量のデータから多くの類義語候補を容易に収集するのに有用である．また２部グラフの一方の頂点集合が類義語グループ，他方がそのグループの観点として解釈できるため類義語のグループ化と同時にそれに付随する観点も生成できる．その結果，多様な観点で類義

† 東京都品川区
Shinagawa-ku, Tokyo, Japan

語のグループを構成することができる．これまで語彙の類似度判別に関する研究としては与えた観点に応じて異なる類似度を計算する方式[7)~9)]が提案されているが，観点自体を生成するものはなかった．さらに本手法はグラフ間の関係によって語彙間の類似度を与えることもできるという特徴をもつ．

以下，第2章では本手法について説明し，第3章では類義語抽出実験とその結果を報告する．第4章では考察を行う．第5章では関連研究について述べる．最後に第6章で今後の課題と展望について述べる．

2. 類義語の抽出手法

ここでは提案する類義語の抽出手法について述べる．

2.1 手法の概要

下記のステップから成る．

1：概念関係抽出　テキストを形態素解析して概念(語彙)間の関係を抽出する．

2：グラフ化　抽出された2項関係を2部グラフに変換する．

3：類義語グループの抽出　この2部グラフから完全2部グラフを抽出する．これが類義語のグループとなる．

4：グラフの選別　完全2部グラフから極大完全2部グラフを選別する．

2.2 類義語抽出の考え方

本節では提案する手法における類義語抽出の基本となる考え方を説明する．語彙間の関係としては名詞の連接した複合語における修飾関係(例：日本経済)，名詞を助詞で結合した関係(例：意見の対立)，文中での主語や動詞といった関係などがあるがここでは例として簡単な複合語の修飾関係を考える．

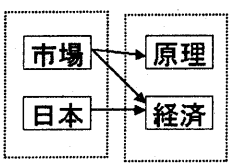

図1　2部グラフ

例えば日本経済という複合語では「日本」が「経済」を修飾している．このとき各語彙を頂点、関係を辺とするようなグラフを複数の複合語{日本経済，市場経済，市場原理}から作成することができる(**図1**)．これをG_1とするとこのグラフは頂点集合{日本，市場}，{原理，経済}からなる2部グラフ☆である．このように連接する名詞からなる複合語の組は，各連接部分の前半と後半の集合を2つの頂点集合(始頂点集合，終頂点集合)とする2部グラフとなる．

このようなグラフの中で特に完全2部グラフ☆☆となるものを考えてみる．例として{中国投資，中国経済，中国市場，日本投資，日本経済，日本市場，米投資，米経済，米市場}から得られるグラフを(**図2**)に示す．これは完全2部グラフである．そしてこのとき始頂点集合V_s={中国，米，日本}，終頂点集合V_e={投資，経済，市場}は類似した意味の語彙グループや同じテーマに属する語彙グループを構成していることがわかる．前者は国名であり，後者は経済活動に関わる語彙である．本手法ではこのような完全2部グラフの頂点集合を類義語グループの候補として抽出するものである．

またこのように抽出されたグループを用いてその中の語彙の類似度を考えることもできる．例えば{a,b,c,d},{p,q,r}という2つの集合からなる2部グラフG_1と{a,b,c},{p,q,r,s}からなる2部グラフG_2を考える．このときV_{e1}より詳しいV_{e2}によって{a,b,c}がグループ化されていると解釈するとV_{s2}の要素間の類似度は{s}という観点のもとでV_{s1}のそれよりも高いと考えることができる．

しかしこのような完全2部グラフであっても頂点集合のサイズ(位数)が小さい場合には類義語グループを構成しない場合もある．例えば次の4語(「世界戦略」，「世界不況」，「長期戦略」，「長期不況」)から導かれるグラフは完全2部グラフ(V_s={世界，長期}，V_e={戦略，不況})であるがV_s，V_eが類義語グループであるとは考えられない．ここでさらに「国際情勢」，「国際戦略」，「国際不況」という語も加えて考えてみると，さらにV_eの大きな完全2部グラフ(V_s={世界，国際}，V_e={戦略，不況，情勢})を得ることができる．そしてこのときV_sは類義語の集合になっている．このように同じ完全2部グラフでも頂点集合の位数が高いほうがより類義語になりやすいことが期待される．

☆ 頂点を2個の頂点集合に分割したとき，全ての辺についてその両方の端点が別々の頂点集合に接続している．

☆☆ 2部グラフの頂点集合をX,Yとするとき X(またはY)の任意の点はY(またはX)の全ての点と接続している

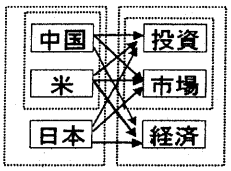

図 2 (極大) 完全2部グラフ

ところで (図 2) にこにおいて頂点集合 { 中国, 米 } と { 経済, 市場 } だけからなるグラフを考えても完全 2 部グラフを構成している. すなわち V_s, V_e が完全 2 部グラフの各頂点集合であるとき任意の $V_s'(\subset V_s)$ と $V_e'(\subset V_e)$ を始頂点集合, 終頂点集合とする部分グラフ G' もまた完全 2 部グラフを構成する. そこで類義語のグループとしてはこのような完全 2 部グラフのうち各頂点集合*が集合の包含関係で極大なものだけを選びこれらのグラフを極大完全 2 部グラフと呼ぶことにする. したがって V,E をそれぞれ頂点と辺の集合とする 2 部グラフ $G = (V_s, V_e, E)$ においてその部分グラフ $G_i = (V_{s_i}, V_{e_i}, E_i)$ が極大完全 2 部グラフであれば任意の $G_j (j \neq i)$ に対して $V_{s_i} \subseteq V_{s_j}$ かつ $V_{e_i} \subseteq V_{e_j}$ ではない.

例として完全 2 部グラフ $K_{2,2} = (V_s, V_e, E)$ **ただし V_s={a,b},V_e={x,y},E={(a,x),(a,y),(b,x),(b,y)} に含まれる全ての部分完全グラフの包含関係を**図 3** に示す. この図で例えば (a)(x,y) は $V_s = \{a\}, V_e = \{x, y\}, E = \{(a, x), (a, y)\}$ なるグラフを表現している. **図 4** はグラフ G_1(図 1) に含まれる全ての部分完全 2 部グラフの包含関係を示している (ただし x=市場,y=日本, a=原理,b=経済). これは図 3 の部分集合になっており, この場合 G_1 における極大完全 2 部グラフは (x)(a,b) と (x,y)(a) の 2 個となる.

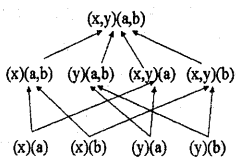

図 3 グラフ間の包含関係

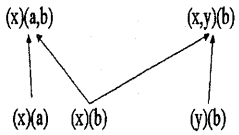

図 4 G_1 の部分完全グラフ間の包含関係

2.3 計算方法

次に与えられた 2 部グラフに含まれる全ての極大完全 2 部グラフを抽出するための計算方法について述べる.

2.3.1 問題の定式化

極大完全 2 部グラフ抽出問題は与えられた 2 部グラフ G に含まれる全ての完全 2 部グラフの中から極大なグラフを抽出する問題である. この全ての完全 2 部グラフを求めるには G の始頂点集合 V_s の部分集合全体すなわち V_s の冪集合 2^{V_s} の要素と終頂点集合 V_e の冪集合 2^{V_e} を比較して, 両者からなる 2 部グラフで完全 2 部グラフを構成するものを見つければよい. その後, 得られたこれら部分グラフの中に $G_1 = (V_{s1}, V_{e1}, E_1), G_2 = (V_{s2}, V_{e2}, E_2), V_{s1} \subset V_{s2}$ または $V_{e1} \subset V_{e2}$ を満たすような G_1 があればこれを除外することによって極大完全 2 部グラフの集合が得られる.

この問題に要する計算量であるが, 部分グラフの抽出だけでも最悪の場合, V_s, V_e の位数 (要素数) を N,M とすると 2^{N+M} 個の部分グラフが存在し, グラフ $K_{i,j}$ には $i \times j$ 本の辺があるから以下の回数だけ辺の接続を調べることが必要である.

$$\sum_{i=1}^{N}\sum_{j=1}^{M} i_N C_i j_M C_j = \sum_{i=1}^{N} i_N C_i \sum_{j=1}^{M} j_M C_j = \frac{NM}{4} 2^{N+M}$$

(二項係数の性質 $N2^{N-1} = \sum_{i=0}^{N} i_N C_i$ より)

しかし現実的には得られる完全 2 部グラフの始頂点集合の位数はある語彙に接続する語彙の種類数が上限なので部分グラフの比較数はこれよりは小さいと予想される.

2.3.2 計算手順

計算の手順を以下に示す. 始頂点集合 V_s, 終頂点集合 V_e, 辺集合 E とする 2 部グラフ $G = (V_s, V_e, E)$ において

(1) 各頂点 $v_i \in V_s$ に接続する頂点の集合を $V_i(\subset V_e)$ を求める.

(2) この V_i から冪集合 2^{V_i} を構成する (その要素を $v_{i_j}, (0 \leq j < 2^{|V_i|})$ とする). このとき $\{v_i\}, v_{i_j}$ で構成するグラフは完全 2 部グラフとなる (これを

---

* 頂点数は 2 以上

** $K_{i,j}$ 2 つの頂点集合の各位数が i,j なる完全 2 部グラフ

G_{i_j} とする). このようなグラフをすべて生成する.

(3) 2で得られた完全2部グラフで $v_{i_j} = v_{p_q}$ となるような2つのグラフ G_{i_j}, G_{p_q} があれば $\{v_i, v_p\}, v_{i_j}$ もまた完全2部グラフとなる. このように完全2部グラフ $G_1 = (V_{s1}, V_{e1}, E_1), G_2 = (V_{s2}, V_{e2}, E_2)$ で $V_{s1} = V_{s2}$ または $V_{e1} = V_{e2}$ であれば V_{s1} と $2^{V_{e1} \cup V_{e2}}$ の任意の要素 (前者の場合. 後者も同様) で2部グラフを構成する. この操作で生成されるグラフはまた完全2部グラフとなる. そこで2で得られたグラフにこの操作を適用し, その結果に対しても順次同様にこの操作を適用して可能な限りグラフを生成する.

以上のステップで得られるグラフ全体が求める部分グラフの集合である.

この際, 上のステップ3の操作で次のグラフ集合を生成できないグラフがあればそのグラフは極大完全2部グラフである. なぜならステップ3の操作で生成される $G_1 = (V_{s1}, V_{e1}, E_1), G_2 = (V_{s2}, V_{e2}, E_2)$ で $V_{s1} = V_{s2}$ のときのグラフ $G_{1+2} = (V_{s1}, V_{e1} \cup V_{e2}, E_1 \cup E_2)$ は G_1, G_2 より大だからである.

3. 実　　験

本手法の有効性を確認するために実験を行った. ここでは実験の詳細について説明する. この実験では抽出される類義語グループの数, 大きさ, その内容が確かに類義語としての集合を形成しているかなどを確認をする. そのためまず実際にテキストデータからグラフを構成し, 次に前述した計算法によって与えられたグラフの全ての部分2部グラフより極大完全2部グラフ (類義語グループ) の抽出を行った.

類義語を抽出するためのテキストデータとしては新聞を用いた. そのデータを形態素解析して得られる概念間の関係としては新聞記事中の複合語における名詞の連接関係を用いた. 新聞は校正水準が非常に高く誤字脱字がほとんど存在しない. また使用語彙に基準があり統制のとれた文章なので品質が安定しているという特徴がある.

3.1 実験資料

テキストデータとして用いた資料について説明する. 資料の特徴を**表1**に示す. テキストデータには佐賀新聞[☆1]の1994年1月の記事3268本 (総文字数2129063文字[☆2]) を使用した. この中に含まれる名詞は約54万語である. これら名詞の連接部分のうち複合語として本実験で使用したは約5.5万語 (選択基準は後述) である. 1日分の新聞記事は次のように1:総合1面, 2:総合その他, 3:国際, 4:経済, 5:地方 (佐賀県, 九州地方の話題が多い), 6:スポーツ, 7:社会, 8:文化, 9:ひろば (読者の投稿など), 10:論説, 11:特集記事 (不定期掲載), 12:死亡 (訃報), 13:情報[☆3]全部で13種類に分類されている. これから分るようにこの分類体系は複数の観点から構成されている. 「総合1面」は注目度の高いあらゆる分野のニュース扱う. 「国際」と「地方」はニュースの発生した場所による分類で「経済」, 「スポーツ」, 「社会」, 「文化」, 「死亡」は内容による分類である. また「ひろば」, 「論説」, 「特集記事」には執筆者がそれぞれ読者, 記者, 依頼されたライターという違いがある.

表1 用いたテキストデータの統計量

掲載月 (1994年)	1
記事数	3268
総文字数	2129063
名詞数 (延べ)	547361
名詞数 (異なり)	36501
複合語数 (延べ)	55567
同 (異なり)	29772

3.2 実験方法

実験手順　実験の手順を2.1節の各ステップに沿って説明する.

- **概念関係抽出**

  形態素解析　この実験においてグラフ化する語彙の関係としては形態素解析により容易に抽出することができ, また語彙相互の結合度も高い関係である複合語に含まれる連接した名詞の関係を用いた. 実際の処理には形態素解析システム「茶筅[☆4]」を用いてテキストデータを形態解析処理し, その結果から名詞だけの連接部分を抽出した. 例えば「政府与党」, 「首脳会議」という名詞の連接は「政府」＋「与党」, 「首脳」＋「会議」に分析されこれにより連接した名詞同士の (政府, 与党), (首脳, 会議) という2項関係が得られる. しかし名詞連接部分には代名詞のように複合語を構成する部分として不適切なものも含まれる. また「二日」という語の場合, 助数詞である「日」は必ず数詞の後に連接することがあらかじめ判明している. グラフ上でも「一日」～「十日」, 「一月」～「十月」から生成

☆1 http://sagashinbun.co.jp/

☆2 元記事に含まれる記号類もそのまま計数している. 改行文字は含まない

☆3 献立のヒント, 記者日記, おくやみ, お誕生日おめでとう等の決まった記事やイベントのお知らせ

☆4 http://chasen.aist-nara.ac.jp

されるグラフから {一, .., 十} と {日, 月} から成る完全2部グラフ (数, 助数詞をグループ化している) が得られることは容易に理解できる. しかも数詞はグラフ中で位数の大きい頂点集合を作るため, これを含む部分グラフの量は極めて多量 (上記の例で $(2^2-1)(2^{10}-1)$ 個) である. そこで今回は連接名詞でもこのような以下のケースを除外した.

- 代名詞 :「これら」等
- 固有名詞-人名
- 数
- 接尾 (語) :「～県」,「～さん」,「～年」,「～化」,「～性」等
- 非自立 :「～もの」,「～ころ」,「午後～」,「～以上」,「従来～」,「すべて～」等

表1に資料から抽出された複合語の延べ数, 異なり数を示す. 名詞の連接した複合語には「首脳/会議」のような1個の連接関係を持つ2個の部分名詞から成る複合語のほかに「CM/製作/会社」のような3個以上の部分名詞を含むものも存在する. この例の場合は「CM-製作」,「製作-会社」という風に直接連接する部分に関係が存在するが「サッカー/日本/代表」のように「サッカー-代表」,「日本-代表」といった直接連接する部分以外に関係がある場合も多く, このような誤りを避けるために今回は2個の名詞による複合語のみを用いた. しかし形態素解析の際の単語境界や品詞の判定の誤りが存在するためにこれでも若干の誤った2項関係が含まれている. **表2** に資料から抽出された複合語に関する統計を示す. これによると最大長は11になる場合 (「那珂川/リバー/サイド/地区/特定/住宅/市街地/総合/整備/促進/事業」) も存在するが, 70%以上は2名詞の複合語である.

表2 複合語の統計量

掲載月 (1994年)	1
総複合語数	29772
2名詞複合語数	22250
同 全体比	0.75
最大複合語長	11

語の計数は異なり数.
複合語長は名詞数が単位.

- **グラフ化**

  2項関係と2部グラフの対応 形態素解析で得られた名詞 (概念) の2項関係 $r=(v_x,v_y)$ の集合を R とするとき任意の $r \in R$ に対して $v_x \in V_s, v_y \in V_e, R=E$ となるグラフ $G=(V_s,V_e,E)$ を対応付ける.

- **類義語グループ抽出**

  極大完全2部グラフの抽出 2.3.2節で述べたように, 与えられたグラフからグラフ集合 $\{G_{i_j}\}$ を構成し, これらを合成しながらそれより大なグラフを生成してゆく. この際, 新たにグラフを生成できないグラフが極大完全2部グラフとなる.

3.3 実験結果

実験の結果について説明する.

3.3.1 形成されるグループの大きさと数

テキストデータから得られた名詞の2項関係22250個を単一のグラフにしたところ $|V_s|^{\star}=7844, |V_e|=6416$ であった. 辺の数は22250であるので V_s, V_e それぞれ1頂点あたり平均2.84, 3.47個の頂点と接続している.

表3 に抽出された極大完全2部グラフの個数をその頂点集合 V の位数 $|V|$ (大きさ) 別に示す. 例えばこの表で行が2, 列が3の要素は抽出された極大完全2部グラフのうち V_s が2個の名詞で V_e が3個の名詞を含むグラフ (例えば {コメ, 食糧} − {不足, 問題, 輸入}) の個数を表している.

表3 実験結果

$\|V_s\|$ \ $\|V_e\|$	2	3	4	5	6	7
2	2075	750	243	96	38	19
3	790	170	25	8	4	2
4	310	24	5	1	0	0
5	110	5	0	0	0	0
6	55	3	0	0	0	0

$\|V_s\|$ \ $\|V_e\|$	8	9	10	11	12	13	16	17
2	9	7	6	8	4	2	1	1

$\|V_s\|$ \ $\|V_e\|$	7	8	9	10	11	12	14	17
2	29	8	4	2	1	2	1	1

3.3.2 抽出内容

抽出された類義語グループの一部を**表4**に示す. ここではグラフのサイズ ($|V_s|$ と $|V_e|$ の組) 毎にグループ化された名詞のサンプルを1件づつ取り上げた (* 印はそのサイズのグラフで唯一の抽出グループ).

抽出された内容を見ると「右, 左」,「女子, 男子」,「小

☆ V_s の要素数 (位数)

さじ，大さじ」,「衆院，参院」(表 4[2,12],[2,17],[8,2]) のような対概念が正確に得られている．また「自民党，社会党」，「高校，大学」，「中国，日本」(表 4[2,10],[2,11],[2,16]) のように同じカテゴリに属する類義語も得られている．さらには以下のように要素数の多い類義語のグループも形成されている．「かたくり粉，しょうゆ，みじん切り，みりん，ゴマ油，砂糖，酒，酢」→調味料,「安治川，九重，駒，八角，武蔵川」→親方,「フィリピン，フランス，メキシコ，ロシア，韓国，米」→大統領制の国家 (表 4[8,2],[5,2],[6,2])

3.4 評　　価

本手法による類義語グループ抽出がどの程度正しく行われたかを評価するために実験結果に対する検査を行った．検査では抽出されたグループの内容を吟味して類義語のグループであるかどうかを人手で確かめた．ただし概念間の関係は多観点による類似性を考えることができるため，評価者によって若干の個人差があることは止むを得ない．この調査ではおよその精度を確認し，間違って抽出されたグループを確認した．その原因については次節で考察を加える．

検査対象となるのはグラフサイズ [2,2] の 2075 個のグループ (この集合を $\mathbb{G}_{[2,2]}$ とする) である．これ以上のサイズのグラフは $\mathbb{G}_{[2,2]}$ の要素だけから合成されるためもし $\mathbb{G}_{[i,j]}(i,j>2)$ の要素であるグループが類義語グループとして不適切ならその原因はその部分グラフである $\mathbb{G}_{[2,2]}$ にも存在するからである．よって誤抽出確認のための検査としては $\mathbb{G}_{[2,2]}$ で十分である．また 2.2 節で述べたように位数の高いグラフでの抽出精度が高くなるのかも検査した．こちらは $\mathbb{G}_{[2,i]}, \mathbb{G}_{[j,2]}(j>6)$ を対象にした．これらの結果を**表 5** に示す．$\mathbb{G}_{[2,2]}$ では約半数のグラフが類義語をグループ化していることが分った．位数の高いグラフの場合の抽出精度は $\mathbb{G}_{[2,i]}(j>6)$ で約 3／4，$\mathbb{G}_{[j,2]}(j>6)$ では約 1／2 であった．この結果は $|V_e|$ の位数が高い時に抽出精度が高いことを示している．

表 5　誤抽出検査結果

	類義語グループであるグラフ数 A	全グラフ数 B	A／B
$\mathbb{G}_{[2,2]}$	1009	2075	0.49
$\mathbb{G}_{[2,i]}(i>6)$	45	57	0.79
$\mathbb{G}_{[j,2]}(j>6)$	22	48	0.46

4. 考　　察

グループ抽出に失敗する原因　実験評価の検査において抽出されたグラフのうち類義語のグループにならなかったケースを調べた．原因としては 1:形態素解析の失敗，2:多義語，3:汎用的な連接語彙 4. 格や属性の違いによるものがある．1 は本手法の原理上，自明なものである．2 の事例としては { チーム，十両 }-{ 編成，優勝 } というグラフがある．これは「十両」が多義語 (車両の意と相撲用語) であることが原因である．しかし一見，類義語に見えない「チーム」と「十両」であるが観点である「編成」に注目するとこれは人の集まりの意である「チーム」と列車の集まりの意である「十両」をグループ化しているとも解釈できる．多義語の扱いは今後の課題である．同様の例では他に「コード」(電線と番号体系の意) がみられた．3 の例では { 運動，基本 }-{ 機能，方針 } が挙げられる．この場合,「基本」が多くの語彙を修飾できるのが原因である．このような語彙はこの他「特別～」,「～関係」,「～関連」,「優良」がある．これらの語彙はそれが連接可能な語彙種類が多いのを利用して除外できると考えられる．4 の例では { 現地，農業 }-{ 研修，生産 } がある．この場合「現地」は動作一般に対してその場所属性を記述している．一方「農業」は「研修」や「生産」の内容を修飾して限定する．このように連接する名詞が非修飾語どの面を修飾するかによってはグループ化できないことがある．

観点の違いによるグループ化　本手法では原理的に観点の違いによる異なったグループ化が可能である．実験結果の中からはそのような例として「テレビ」が見られた．これは以下のようなグループを形成している．

1 : (テレビ，報道) – (各社，番組)
2 : (テレビ，ニュース) – (画面，番組)
3 : (スポーツ，テレビ) – (観戦，小説)
4 : (テレビ，パソコン) – (ゲーム，画面)
5 : (テレビ，ハイビジョン) – (中継，番組)
6 : (テレビ，フジテレビ) – (社長，番組)
7 : (テレビ，佐賀新聞社) – (社長，編集)

この結果は「テレビ」が多様な観点によって他の類似概念とグループ化できることを示している．これらはそれぞれ「テレビ」が番組 (1,2) や娯楽 (3,4) としての側面を持つこと，上位技術 (ハイビジョン) との対比があること，またこの語が場合によって「機器 (4)」「放送局 (6)」「マスコミ (7)」の意で用いられていることを分りやすくする．従ってこの語が多義語であることを知る手がかりにもなる．

語彙の類似度　2.2 節で複数グループの関係から語彙の類似度を与える方法を示した．ここでは実験結果

表 4 抽出されたグループの例

$[\|V_s\|, \|V_e\|]$	グループ化された名詞
[2, 2]	(環境:人権)-(保護:問題)
[2, 3]	(事故:地震)-(情報:発生:被害)
[3, 2]	(オーストラリア:フランス:日本)-(政府:大使館)
[2, 4]	(フランス:ロシア)-(外相:革命:政府:大統領)
[3, 3]	(オーストラリア:中国:日本)-(産米:政府:米)
[4, 2]	(ウクライナ:シリア:フランス:ロシア)-(外相:大統領)
[2, 5]	(建築:土木)-(技術:工事:談合:畑:部門)
[3, 4]	(経済:地域:農業)-(活性:社会:政策:問題)
[4, 3]	(経済:社会:就職:政治)-(活動:状況:問題)
[5, 2]	(安治川:九重:駒:八角:武蔵川)-(親方:部屋)
[2, 6]	(参院:衆院)-(議員:議運委:議長:採決:事務:段階)
[3, 5]	(伊万里:佐賀:鹿島)-(市長:市内:市役所:地区:保健所)
[4, 4]	(伊万里:佐賀:鹿島:鳥栖)-(市長:市内:地区:保健所)
[5, 3]	(アジア:欧州:世界:中国:日本)-(経済:最大:市場)
[6, 2]	(フィリピン:フランス:メキシコ:ロシア:韓国:米)-(政府:大統領)
[2, 7]	(市:町)-(勤労:社協:助役:職員:体協:中心:農業)
[3, 6]	(佐賀:鹿島:鳥栖)-(市教委:市長:市内:支部:地区:保健所)
[4, 5]*	(国内:中国:日本:米国)-(メーカー:各地:企業:経済:市場)
[6, 3]	(コメ:環境:経済:農業:福祉:流通)-(政策:対策:問題)
[2, 8]	(経済:雇用)-(安定:環境:計画:情勢:状況:審議:対策:問題)
[2, 9]	(経済:行政)-(システム:運営:改革:関係:経験:担当:長官:政策:問題)
[2, 10]	(自民党:社会党)-(幹部:議員:市議:自身:執行:首脳:席:大会:抜き:分裂)
[2, 11]	(高校:大学)-(チーム:講師:最後:時代:受験:生活:選手権:卒業:日本一:入学:入試)
[2, 12]	(右:左)-(CK:ひざ:オーブン:カーブ:サイド:ラインアウト:下手:胸:四つ:手首:太もも:半身)
[2, 13]	(自民党:党)-(改革:幹事:幹部:関係:議員:建設:執行:首脳:出身:大会:抜き:分裂:本部)
[2, 16]*	(中国:日本)-(メーカー:各地:企業:経済:国内:国民:最古:最大:産米:市場:進出:政府:全土:舞踊:文化:米)
[2, 17]*	(女子:男子)-(シングル:シングルス:スピード:ダブルス:テニス:ベスト:回転:学生:決勝:私立:準決勝:小学生:生徒:選手:総合:団体:中学生)
[7, 2]	(伊万里:京都:佐賀:鹿島:大阪:鳥栖:唐津)-(市内:地区)
[8, 2]	(かたくり粉:しょうゆ:みじん切り:みりん:ゴマ油:砂糖:酒:酢)-(小さじ:大さじ)
[9, 2]	(医療:教育:交通:行政:国際:政府:捜査:日本:報道)-(関係:機関)
[10, 2]	(強化:経済:国連:支援:社会:就職:政治:地域:犯罪:保護)-(活動:問題)
[11, 2]*	(アジア:欧州:国際:国内:世界:中国:東南アジア:統合:日本:米:米国)-(経済:市場)
[12, 2]	(コメ:開放:外交:環境:経済:国連:財政:地域:農業:福祉:貿易:流通)-(政策:問題)
[14, 2]*	(プロ:外国:強豪:県:出場:女子:招待:人気:全日本:相手:代表:日本:優勝:有力)-(チーム:選手)
[17, 2]*	(ごみ:エイズ:コメ:ボスニア:環境:経済:雇用:国内:財源:支援:社会党:農業:犯罪:福祉:保護:暴力団:流通)-(対策:問題)

*印は $[x, y]$ の大きさのグラフにおける唯一の抽出結果

より実例を用いて説明する. 国名がグループ化されており含まれる語彙が似ている以下の 3 件のグラフを考える.

G_1:(フランス, ロシア)–(外相, 革命, 政府, 大統領)

G_2:(ウクライナ, シリア, フランス, ロシア)–(外相, 大統領)

G_3:(フィリピン, フランス, メキシコ, ロシア, 韓国, 米)–(政府, 大統領)

G_1, G_3 を比較すると $V_{s1} \subset V_{s3}, V_{e3} \subset V_{e1}$ であるからフランス, ロシア間の類似度は { 外相, 革命 } という観点のもとで, その他のフィリピン, メキシコ, 韓国, 米とのそれより高いといえる. これをフランス基点とした順序関係で示すと, フランス<ロシア<フィリピン, メキシコ, 韓国, 米となる. 同様に観点 { 外相 } ではウクライナ<シリア<フィリピン, メキシコ, 韓国, 米 がいえる. この状態ではフランスから見てフィリピン, メキシコ, 韓国, 米は等距離になる. しかし新たな語彙の関係を追加することにより順序関係は変化する. 例えば (メキシコ)-(外相) を加えると G_2 は極大グラフではなくなる. 代わって G_2':(ウクライナ, シリア, フランス, メキシコ, ロシア)–(外相, 大統領) が新たな極大グラフとなり, 順序関係も フランス<ロシア<メキシコ<フィリピン, 韓国, 米 に変化してより分離度の高い順序が得られる.

5. 関連研究

ここでは関連する研究との比較を行う. 従来, 伝統的な木構造の類義語辞書には分類語彙表[1] や日本語語彙体系[2] などがあった. しかしこれらを用いて言語処理に要求されるような多観点による類似性判別をすることはできない. これは文献 3) にもあるように編集方針でもある. そこで文献 4) は多観点シソーラスを

提案している．この研究では固定した多観点の体系を提案し観点による類似性判別機能を実現した．しかし体系に属する語彙集合の自動収集方法は扱っていない．本手法は類義語を抽出する方法を提案している点が異なる．一方，辞書やコーパスから語彙知識を獲得する研究[5),6)]があるがその結果に基づいて類似性判別までを行う方式は文献 7)~9) が提案している．これらの基本的なアイデアは文脈や語彙の属性を複数の単語で表し，複数単語のベクトル空間上での距離計算によって類似度を計算するものである．距離計算の方法に工夫があるが，類似度自体は数値であるため，外から与えた観点に応じてその値が変化するようになっている．これに対して本手法はある語彙を含む類義語グループを複数生成し，それぞれに異なった観点を付与することができる．つまり観点も生成できる点が異なる．

6. おわりに

本論文ではテキスト中の語彙の関係から 2 部グラフを構築し、そこから極大完全 2 部グラフを抽出することにより類義語グループを獲得する方法を提案した．また語彙間の関係として 2 単語複合語から得られる連接関係を用いた類義語グループ抽出実験を行い，抽出されたグループの約半数が実際に類義語をグループ化していることを確認した．

本手法は

- 大量のデータから多くの類義語候補を容易に収集するのに有用である．
- 抽出された 2 部グラフの一方の頂点集合が類義語グループであるとき，他方はその観点として解釈できるため類義語グループと同時に多様な観点も生成することができる．
- グラフの頂点集合間の関係によって語彙の類似度を定義できる．
- 類義語のグループは語彙間の関係から生成される完全 2 部グラフに基づいているため，抽出された原因とそのグループの意味が直感的に理解しやすい．

といった特徴を持つ．

本手法の課題としては語彙間関係の種別に応じたグラフ構造の改善，多義語の適切な扱い方の工夫などがあげられる．今後は実際のシソーラス構築に本手法を適用してその有用性の評価を行う予定である．

参考文献

1) 国立国語研究所 (編):分類語彙表，秀英出版 (1994).
2) 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 (編): 日本語語彙体系, 岩波書店 (1997).
3) 中野 洋:分類語彙表の増補とその利用, 言語処理学会第 1 回年次大会, pp. 141–144 (1995).
4) 川村 和美, 片桐 康裕, 宮崎 正弘:語を種々の観点から分類した多次元シソーラス, 信学技報, Vol.NLC94-48, pp. 33–40 (1995).
5) 新納 浩幸:コーパスを利用した分類語彙表の未登録語義の発見, 情報処理学会論文誌, Vol.38, No.5, pp. 953–961 (1997).
6) 川前 徳章, 青木 輝勝, 安田 浩:統計的モデルを用いた単語クラスタリング, 情報処理学会研究報告, Vol.NL144-8, pp. 55–60 (2001).
7) 北川高嗣, 清木 康:意味の数学モデルとその実現方式について, 信学技報, Vol.DE93-4, pp. 25–31 (1993).
8) 笠原 要, 松澤和光, 石川 勉, 河岡 司:観点に基づく概念間の類似性判別, 情報処理学会論文誌, Vol.35, No.3, pp. 505–509 (1994).
9) 笠原 要, 松澤和光, 石川 勉:概念知識の構築と判別, 情報処理学会論文誌, Vol.38, No.7, pp. 1272–1283 (1997).