2部グラフ構造を用いた類義語の抽出

○中渡瀬　秀一

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

○ HIDEKAZU NAKAWATASE

Abstract

This paper proposes a method for similar words groups extraction from nouns in a Japanese 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 maximal complete. Further, this method can give also similarity measurements between words using relations among those similar words groups.

1. はじめに

本稿では語彙間の関係の集合から導出される2部グラフの構造的特徴を利用して，自動的に類義語のグループを抽出する方法を提案する．

情報検索などのシステムにおいては，語彙間の類似度計算や，ある語彙に対する類似語彙の獲得が要求される．そこでこのために類義語辞書が今まで用いられてきた．しかしこの構築はほとんど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．特にこの分類体系を作成するために，必要なカテゴリとそれらの構造を定めるのは難しい．従来の類義語辞書[1),2)]では木構造，上位下位関係等がよく使用されてきた．ところが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[3)]．

そこで筆者[7)]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を提案する．提案方法ではテキストを形態素解析して語彙間の関係を獲得し，これを2部グラフに変換する．このグラフの構造的特徴から類義語グループとそのグループの観点を自動抽出する．この方法によって大量のデータから類義語候補を容易に収集可能である．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[4)~6)]など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章では本手法を説明し，第3章で類義語抽出実験結果を報告する．第4章では考察を行う．最後に第5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

本節では提案する類義語の抽出手法について述べる．

2.1 類義語抽出の考え方

まず類義語抽出の基本となる考え方を例で説明する．語彙間の関係として複合語における修飾関係 (例:日本経済) を考えてみる．「日本経済」では「日本」が「経済」を修飾している．ここで各語彙を頂点，関係を辺に対応付けたグラフを考える．このようなグラフ化を複数の複合語 { 日本経済, 市場経済, 市場原理 } に行い，併合すると図1のようになる．このグラフは頂点集合 { 日本, 市場 }，{ 原理, 経済 } からなる2部グラフという特徴がある (グラフの頂点を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含まれる頂点に接続している)．2部グラフの中で特に完全2部グラフ (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) となるものを考えてみる．その例として { 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．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．

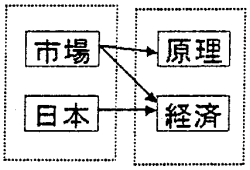

図 1　2部グラフ

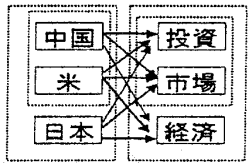

図 2　(極大) 完全 2 部グラフ

しかし完全 2 部グラフにおいては，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 (頂点数は 2 以上) が集合の包含関係で極大なもの (極大完全 2 部グラフ) だけを選ぶ．

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 の部分集合にもなっており，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．

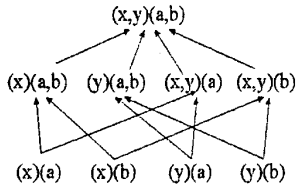

図 3　$K_{2,2}$ の部分グラフ間の包含関係

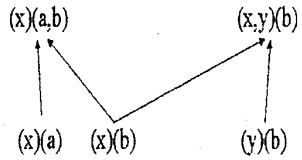

図 4　G_1 の部分完全グラフ間の包含関係

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 2 つの完全 2 部グラフ (図 4 の記法と同様) $G_1 = (a, b, c, d)(p, q, r)$ と $G_2 = (a, b, c)(p, q, r, s)$ を考える．このとき (a, b, c) は (p, q, r) より詳しい (要素の多い)(p, q, r, s) によってグループ化されていると考えると，この要素間の類似度は {s} という観点のもとで (a, b, c, d) のそれよりも高いと考えることができる．すなわち a,b の類似度を d(a,b) とすると G_1, G_2 においても，d(a,b) = d(b,c) = d(c,a) であるが G_2 における類似度を優先して，$d(a, b) < d(a, d)(= d(b, d), d(c, d))$ という類似度の順序関係が得られる．

2.2　手法の概要

抽出手法をまとめると以下のようになる．

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 2 項関係を抽出する．

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 部グラフから極大完全 2 部グラフを抽出する．

2.3　計 算 方 法

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 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．

表 1 用いたテキストデータ (佐賀新聞 1994 年 1 月) の統計量

記事数	総文字数	名詞数 (延べ)	左同 (異なり)	複合語数 (延べ)	左同 (異なり)	2 名詞複合語数	同全体比	最大複合語長
3268	2129063	547361	36501	55567	29772	22250	0.75	11 語

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．実験ではまず前述の手順により実際のテキストから極大完全 2 部グラフ (類義語グループ) の抽出を行った．次に抽出される類義語グループの数，大きさ，その内容の確認を行った．用いたテキストデータは新聞 (校正水準が高く形式が安定しているので本手法に望ましい) である．グラフ化する概念間の関係は記事中の複合語における名詞の連接関係から抽出した．

3.1 実 験 資 料

用いたテキストデータについて説明する (表 1)．テキストデータは佐賀新聞の 1994 年 1 月の記事 3268 本である．この中に名詞は約 54 万語含まれていた．さらに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 万語である．1 日分の新聞記事は 1:総合 1 面，2:総合その他，3:国際，4:経済，5:地方，6:スポーツ，7:社会，8:文化，9:ひろば，10:論説，11:特集記事，12:死亡，13:情報の全部で 13 種類に分類されている．

3.2 実 験 方 法

実験手順　実験の手順を 2.2 節の各ステップに沿って説明する．

- **概念関係抽出**　グラフ化に用いる語彙の関係は複合語に含まれる連接した名詞の関係である．これらは形態素解析ツールにより容易に抽出可能である．実際の処理には形態素解析システム「茶筅」(http://chasen.aist-nara.ac.jp) を用いた解析結果から名詞だけの連接部分を抽出した (例:「政府与党」という複合語を「政府」+「与党」に分解し，[政府, 与党] という 2 項関係を得る)．しかし名詞の連接がすべて複合語を構成するわけではない (例えば代名詞を含む場合など)．また助数詞 (「月」など) を含む場合，これより極めて多数のグラフが生成されるため今回は連接名詞でもこのような語である代名詞，固有名詞人名，数，接尾 (語)，非自立 (語) を含むものを除外した．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」のように 2 語からなるものの他,「ＣＭ/製作/会社」のような 3 個以上の名詞を含むものもあるが，その場合，直接連接する部分に概念関係ない場合もあるので，今回は 2 個の名詞による複合語のみを用いた．表 1 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 70%以上は 2 名詞の複合語である．
- **グラフ化**　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 $v_x \in V_s, v_y \in V_e, E = R$ 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**　2.3 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 $|V_s| = 7844, |V_e| = 6416$ であった．**表 2** 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 $|V|$ (サイズ) 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の個数を表している．

3.3.2 抽 出 内 容

抽出された類義語グループの一部を**表 3** に示す．抽出された内容を見ると「右，左」,「女子，男子」,「小さじ，大さじ」,「衆院，参院」(表 3[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

表 2 実験結果

		2	3	4	5	6	7	8	9	10	11	12	13	16	17	$\|V_e\|$
	2	2075	750	243	96	38	19	9	7	6	8	4	2	1	1	
	3	790	170	25	8	4	2									
$\|V_e\|$	4	310	24	5	1	0	0									
	5	110	5	0	0	0	0									
	6	55	3	0	0	0	0									
		7	8	9	10	11	12	14	17	$\|V_s\|$						
$\|V_e\|$	2	29	8	4	2	1	2	1	1							

表 3 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:オープン: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

党」,「高校，大学」,「中国，日本」(表 3[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 3[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精度の評価を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人手で確かめ (判定基準は後述), 適合率を計算した．ただし評価者によって判定の個人差が若干存在す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の内容を確認した．その原因については 4 節で考察を加える．類義語の判定に用いた基準について述べる．今回は 3 種類

の関係 (同義：ほとんど同じ意味, 同類：同じカテゴリに属する, 上位下位：属するカテゴリに順序が存在) にある語彙同士を類義語とした．表 4 に検査結果 (適合率) を示す．この表も表 2 と同様に V の位数別に整理している．この結果をみると位数の高いグラフでの適合率が高くなっていることが分る．特に $|V_s| = |V_e| = 2, |V_e| = 2$ の場合を除けば $|V_s|$ または $|V_e|$ が一定のとき，$|V_e|$ または $|V_s|$ が増加するほど適合率は高くなっていることが分った．

表 4　誤抽出検査結果 (適合率)

		2	3	4	5	6	7	8	9	10	11	12	13	16	17	$\|V_e\|$
	2	0.29	0.21	0.23	0.42	0.42	0.47	0.78	0.71	0.67	0.88	0.75	1.00	1.0	1.0	
	3	0.16	0.22	0.44	1.00	0.75	1.00									
$\|V_s\|$	4	0.16	0.25	1.00	1.00	-	-									
	5	0.12	0.80	-	-	-	-									
	6	0.25	0.67	-	-	-	-									

		7	8	9	10	11	12	14	17	$\|V_s\|$
$\|V_e\|$	2	0.34	0.25	0.00	0.50	0.00	0.50	1.0	0.0	

4. 考　　察

グループ抽出に失敗する原因　実験結果の検査を通して，正しく抽出できない原因について調べた．これには 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．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．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」を修飾してその目的や対象を限定する．このように連接する名詞が非修飾語の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　本手法は観点の違いによる異なったグループ化が可能である．実験結果ではこのような例として次に示すような「テレビ」を含むグループが見られた．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．つまり「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　2.1 節で語彙の類似度を与える方法を示した．ここでは実験結果の実例を用いて説明する．次の 3 個のグラフを考える．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．

5. 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．この手法では類義語グループと同時に多様な観点も生成することが可能であり，グラフの頂点集合間の関係によって語彙の類似度を考えることもできる．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

10%($|V_s|=5, |V_e|=2$)〜100%($|V_s|=4, |V_e|=4$) 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, 計算手法の改善による計算量の低減などがあげられる.

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) 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).
4) 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).
5) 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).
6) 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).
7) 中渡瀬秀一:複合語からの類義語抽出法, 情報処理学会研究報告, Vol.FI66-6(DD32-6), pp. 39–46 (2002).