辞書にもとづいて語彙をクラスタリングする試み

小嶋 秀樹　　伊藤 昭
郵政省 通信総合研究所 関西先端研究センター

本論文では，与えられた単語集合 K を手がかりとして，K と意味的に関連するクラスタ（語彙 V の部分集合）C を生成する手法を提案する．本手法では，手がかり K の意味的な「方向性」を抽出し，同じ方向性をもつ単語集合 C を語彙 V から切り出す．まず，英語辞書から構成した意味ネットワーク上の活性伝播によって，各見出し語の意味ベクトル（2851次元）を生成し，この意味ベクトルについて主成分分析を行なう．つぎに，得られた主成分軸のなかで，手がかり K の分布が偏っているものを取り出し，それらの同時分布からクラスタ C を得る．この手法は，(1) 語彙 V を直和分割するのではない点と，(2) 手がかり K の「方向性」を考慮する点で，従来のクラスタリング手法とは異なっている．

1 はじめに

単語間（または文やテキストの間など）の意味的な「類似性」を捉えることが，自然言語処理の多くの分野で必要とされている．類似性にもとづく手法は，情報検索やテキスト構造の認識などに有効である.[4] また，コーパス処理においても，類似性にもとづいて語彙をクラスタリングすることによって，コーパスのスパース性に対処できることが示唆されている.[1]

単語間の意味的な類似性を捉える方法として，コーパスにおける単語の共起性から抽出する方法（[2] など）と，辞書の語義定義から抽出する方法（[3] など）があげられる．本研究では，辞書にもとづく方法をとる．その理由は，十分大きな語彙について類似性を捉えるには，本質的にスパースなコーパスにたよることなく，単語の意味を直接扱うことが必要だからである．むしろ，本論文で提案する類似性の計算方法は，コーパス処理におけるスパース性の問題に対処するための手段となる．

従来からある単語の類似性やクラスタリングの研究[1,2,3] では，単語間の類似度を静的（文脈に関係なく一定）なものとして扱うことが主流であった．しかし，単語間の類似度は，文脈や注意の「方向性」に依存して動的に変化する．たとえば car に類似した単語として，bus, truck, motorcycle などが連想されることもあれば，engine, tyre, seat などが連想されることもある．たとえ文脈なしの自由連想でも，人間は何らかの方向性を想定して，類似した単語を連想していると思われる．

単語間の類似度を計算するには，文脈や注意の方向性が与えられなければならない．そこで本研究では，従来のクラスタリング問題「語彙 V のクラス $\{C_i\}$ への直和分割」ではなく，

> 「与えられた手がかり K から，文脈や注意の方向性を考慮して，類似した単語のクラスタ（語彙 V の部分集合）C を生成する」

という問題を扱う．「文脈や注意の方向性」は，手がかり K を単語集合とし，そこから抽出することとする．たとえば，$K=\{\text{apple, grape, orange}\}$ から *fruit* という方向性を抽出し，$C=\{\text{apple, banana, cherry, } \cdots\}$ を生成するというものである．

以下，第2節では，英語辞書から構成した意味ネットワークによって，語彙 V の各単語を意味ベクトルに変換することを説明し，第3節では，この意味ベクトルを主成分分析することによって，互いに直交した主成分軸を取り出すことを説明する．第4節では，手がかり K の「方向性」を抽出し，目的とするクラスタ C を生成する方法を説明する．この方法は，主成分軸のなかから，手がかり K の分布が偏っているものを取り出し，これらの同時分布としてクラスタ C を得るというものである．第5節では，ここで提案した手法について，認知的な側面から考察する．最後の第6節で，本論文をまとめ，今後の課題を述べる．

2 辞書を利用した意味ベクトルの生成

まず，語彙 V の各単語を意味ベクトルに変換する．この変換には，英語辞書 LDOCE (*Longman Dictionary*

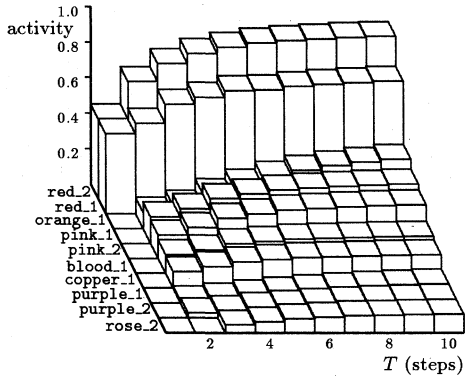

図1 意味ネットワークによる意味ベクトルの生成例
(見出し語 $w=$ red を活性化させたとき, 活性度上位 10 節点の活性度分布を記録したもの. $T=10$ の活性度分布を活性パターン $P(w)$ とする.)

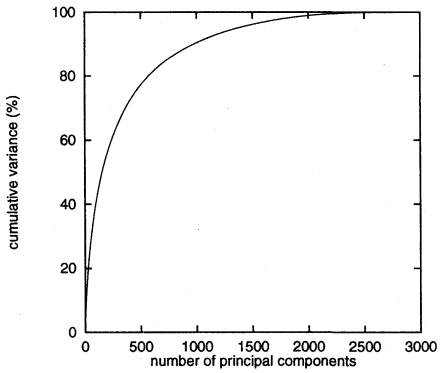

図3 主成分軸上の意味ベクトルの累積分散
(主成分軸 $S_1,\cdots,S_k$ について, 意味ベクトルの累積分散 $\sum_{i=1,k} v_k$ をグラフにしたもの.)

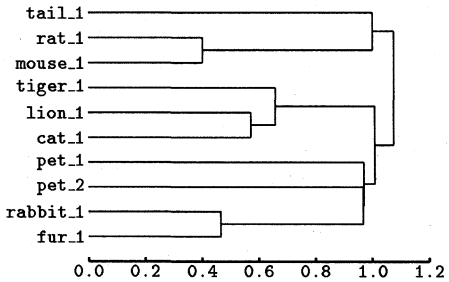

図2 意味ベクトルの階層クラスタリング
(ユークリッド距離 (重心法) によるデンドログラムの一部分. 方向性の異なる関連語 (tail, fur) が含まれている.)

of Contemporary English) から機械的に構成した意味ネットワーク *Paradigme* [3] を利用する. また, ここで扱う語彙 V は, LDOCE の定義用語彙 LDV (2851 語) とする. *Paradigme* は, V (=LDV) の各単語に対応する 2851 の節点と, それらの LDOCE における語義定義に対応する 295914 のリンクからなる. ある節点 w を活性化させることによって, その活性がリンクをとおして伝播し, *Paradigme* 上に活性パターン $P(w)$ をつくりだす (**図1**). 本論文では, この活性パターン $P(w)$ を見出し語 w の意味ベクトル (2851 次元) とする.

Paradigme によって生成された意味ベクトルの特徴をつかむため, $\langle w,w'\rangle \in V^2$ についてユークリッド距離 $d(P(w),P(w'))$ を計算し, 階層クラスタリング (重心法) を行なった結果を **図2** にあげる. 直観に合った類似性 (rat–mouse, tigar–lion–cat) もあるが, 性質の異なる単語 (他の動物にも関係する tail, fur) が入り込んでいることがわかる. この結果から, 人間の直観による類似性に近づくためには, 意味ベクトル間の類似度に「方向性」を与えることが必要であることがわかる.

3 意味ベクトルの主成分分析

つぎに, 前節で生成した意味ベクトル $P(w_1),\cdots,P(w_{2851})$ を主成分分析する. すなわち, 意味空間 R^{2851} を 1 次元直交部分空間 $S_1,\cdots,S_{2851}$ に分解し, それぞれの S_i (これを第 i 主成分軸とよぶ) に意味ベクトルを射影する. このとき, 意味ベクトルを S_1 に射影したときの分散 v_1 が最大となり, S_2 に射影したときの分散 v_2 が 2 番目 $\cdots$, となるようにする. 言い換えれば, S_1 は最も大きな情報量をもち, S_2 はつぎに大きな情報量をもつ.

主成分分析の結果, 意味ベクトルを主成分軸 S_i に射影したときの分散 v_i は, つぎのようになった.

i	v_i
1	0.01922566
2	0.00786715
3	0.00683586
4	0.00571597
5	0.00536231
⋮	⋮

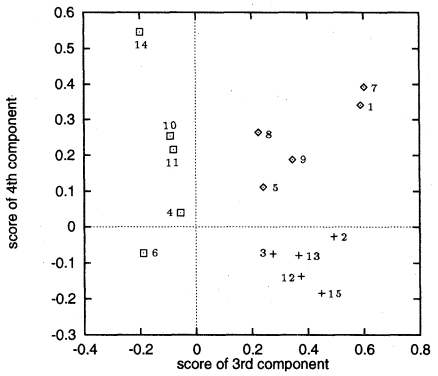

図4 意味空間での単語の分布
(主成分軸 S_3, S_4 によって張られる平面上に, 本文中にあげた 15 単語を散布したもの. 「楽しみ」 (◇), 「苦しみ」 (+), 「結婚」 (□) に関する単語が偏りをみせる.)

また, v_i の累積 $\sum_{k=1,i} v_k$ をグラフにすると**図3**のようになった. 主成分軸は全部で 2851 個あるが, 最初の 10 個だけで全情報の 10.73% を表現でき, 100 個なら 40.67%, 500 個なら 77.16%, 1000 個なら 90.41% を表現できることがわかる.

いくつかの主成分軸によって張られる意味空間の性質をつかむため, つぎにあげる単語

$$\begin{aligned}&\text{amusing}_1,\ \text{angry}_2,\ \text{anxiety}_3,\ \text{ceremony}_4,\\ &\text{entertain}_5,\ \text{faithful}_6,\ \text{funny}_7,\ \text{humorous}_8,\\ &\text{laughter}_9,\ \text{marriage}_{10},\ \text{marry}_{11},\ \text{offend}_{12},\\ &\text{severe}_{13},\ \text{unite}_{14},\ \text{upset}_{15}\end{aligned}$$

の意味ベクトルを, 主成分軸 S_3, S_4 によって張られる部分意味空間 (平面) に射影した. **図4**に示すように, 「楽しみ」に関する単語 (◇) は右上に, 「苦しみ」 (+) は右下に, 「結婚」 (□) は左側に, それぞれ分布している. また, S_3 だけでは「楽しみ」と「苦しみ」を区別できないことも読みとれる. このように,

> 「意味的に類似した単語のクラスタは, いくつかの主成分軸上で偏った分布をもつ」

ということがわかる.

4 方向性の抽出とクラスタ生成

手がかりとして与えられた単語集合 K から「方向性」を抽出することは, いくつかの主成分軸によって K の分布を捉えることである. つまり, 意味空間における K の分布を効率よく説明できるいくつかの主成分軸 $S'_1, \cdots, S'_n \in \{S_i\}$ と, 各 S'_i 上での K の分布特徴を見つければよい. たとえば, S_3, S_4 に注目したばあい,

$$\begin{aligned}K = \{&\text{funny, humorous, laughter}\}\\ &\text{funny} \quad \mapsto \quad (0.60211,\ 0.39212)\\ &\text{humorous} \quad \mapsto \quad (0.22377,\ 0.26478)\\ &\text{entertain} \quad \mapsto \quad (0.24068,\ 0.11190)\end{aligned}$$

の分布特徴は, 2 つの実数集合

$$\begin{aligned}&\{0.22377,\ 0.24068,\ 0.60211\}\\ &\{0.11190,\ 0.26478,\ 0.39212\}\end{aligned}$$

のそれぞれの分布特徴として抽出される.

では実際に, 主成分軸を選び出すにはどうすればよいだろうか. また, そこからどのような分布特徴を取り出せばよいだろうか. いろいろな選択基準や分布モデルが考えられるが, 本論文では, つぎにあげる「区間被覆モデル」を考える. すなわち, 各主成分軸 $S \in \{S_i\}$ について, 以下の操作を行なう.

1. 語彙 V の各単語 w の意味ベクトル $P(w)$ を S 上に射影し, その最小値 A と最大値 B を得る. また, $R = A - B$ とする.
2. 手がかり K の各単語 k の意味ベクトル $P(k)$ を S 上に射影し, その最小値 a と最大値 b を得る. また, $r = a - b$ とする.
3. 偏りの度合い $f = r/R$ を求める. この f がある閾値より大きいときは, 主成分軸 S をすてる. そうでなければ, S における K の拡大分布区間を $I = [a - \mu r, b + \mu r]$ とする.

ただし, μ は分布区間を拡大するための係数で,

$$\mu = \frac{1}{2(|K|-1)}$$

と定義する. この手続きによって, 主成分軸 $S'_1, \cdots, S'_n$ が選び出され, その上での拡大分布区間 $I_1, \cdots, I_n$ が得られる.

目的とする単語クラスタ (または意味空間の部分集合) は, 手がかり K から上述の手続きによって得られた分布区間 $I_1, \cdots, I_n$ の同時分布である. すなわち, 選び出された主成分軸 $S'_1, \cdots, S'_n$ によって張られる意味空間のなかで, 手がかり K の各単語を (ある程度の余裕をもって) 内包するような超直方体が得られる (**図5**). この超直方体の内部に写像される単語の集合が, 目的とする単語クラスタ C である.

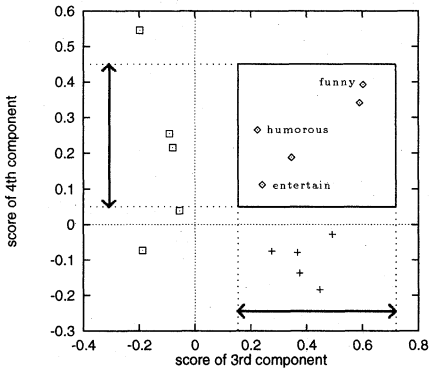

図5 意味空間からの超直方体の切り出し
(手がかり K の主成分軸 S_3, S_4 上での分布区間にもとづいて, 意味空間 $S_3 \times S_4$ から長方形 (超直方体) を切り出す. 図は $K=\{\text{funny, humorous, laughter}\}$ のばあい.)

5 認知的側面の考察

人間にとくに優れた能力として,「情報の重要な部分に注目し, ほかの部分には注意を払わない」という能力と「刻々と変化する環境に応じて, 注意の方向性を変えていく」という能力がある. このような「文脈」を捉える能力によって, 人間は「フレーム問題」を回避することができ, 少ないコストで準最適な解を求めることができる. しかし, 人間のもつこのような能力は, 計算機の最も不得意とする領域であり, 人工知能システムやインタフェースが欠陥を露呈する穴でもある.

本論文で提案した手法 — 少数の主成分軸に注目し, 手がかり K の「方向性」を捉えること — は, このような能力の一部分を定式化したものといえる. また, 手がかり K を文脈に応じて動的に変化させ, その方向性をトレースすることも考えられる. たとえば, 手がかり K を短期記憶 (いわゆる 7 ± 2 チャンク) の内容とみなすことによって, 環境の変化に自律的に適応していく能力をモデル化できるかもしれない.

6 おわりに

本論文では, 与えられた手がかり K から, その意味的な「方向性」を考慮して, 関連する単語のクラスタ C を求める方法を提案した. K の方向性は, 主成分分析によって直交化された意味空間における K の分布の偏りとして抽出される. この分布の偏りにもとづいて部分意味空間が取り出され, 目的とするクラスタ C が得られる.

提案した単語クラスタ構成法は, (1) 語彙 V を直和分割するのではなく, 与えられた手がかり K にもとづいて動的に単語クラスタを構成する点と, (2) 手がかり K の「方向性」を考慮することによって, 文脈に依存した類似性を捉えられる点で, 従来のクラスタリング手法とは異なっている. 本方法は, 状況・履歴に応じて動的に変化するシソーラスを可能にし, コーパス処理におけるスパース性への対応や, より柔軟な自然言語処理などに応用できるものである.

今後の課題として, まず (1) 主成分軸上での K の偏りをより効果的に抽出することと, (2) 主成分軸の選択基準を見直すことがあげられる. 本論文では単純な「区間被覆モデル」を使ったが, このままでは手がかり単語の数 $|K|$ につれて, 得られる単語クラスタの大きさ $|C|$ は増加する傾向をもつ. 何らかの分布モデルを持ち込むことやノイズ (外れ値) への対応などが必要であろう. また, 現在のところ本手法の基本的な枠組みができたところであり, まだ大規模な評価実験は行なっていない. 今後は (3) 実験をとおして人間の直感と比較したり, 言語処理への応用をとおして評価を行なうことが必要である.

参考文献

[1] P. F. Brown, *et al.* : Class-based n-gram models of natural language, *Computational Linguistics*, Vol.18, pp.467–479, 1992.

[2] K. W. Church and P. Hanks : Word association norms, mutual information, and lexicography, *Computational Linguistics*, Vol.16, pp.22–29, 1990.

[3] H. Kozima and T. Furugori : Similarity between words computed by spreading activation on an English dictionary, in *Proceedings of EACL-93*, pp.232–239, 1993.

[4] H. Kozima and T. Furugori : Segmenting narrative text into coherent scenes, *Literary and Linguistic Computing*, Vol.9, pp.13–19, 1994.