Webページを用いた基底単語選出

藤井丈明
茨城大学大学院
理工学研究科

新納浩幸
茨城大学工学部
システム工学科

佐々木稔
茨城大学工学部
情報工学科

1 はじめに

本研究では，Webページを用いた基底単語選出を行う．自然言語処理において，単語に表層的な特徴抽出を行い個々の特徴の重みを表す数値を要素とする，つまり特徴ベクトル化することは非常に重要である．単語を特徴ベクトル化する際には，通常基底の単語が用いられている．基底の単語とは，共起頻度を得るための指標となる単語のことである．基底の単語と特徴ベクトル化する単語との，コーパス内における共起頻度を用いて特徴ベクトルの作成を行うが，新聞記事等をコーパスに用いて特徴ベクトルを作成した場合，情報量が制限されるため，スパース性の問題が発生する．スパース性を解消するために，Webをコーパスに用いて特徴ベクトルを作成したが，最適な特徴ベクトルが得られなかったことが報告されている [1]．これは，Webをコーパスに用いた際の基底の単語に問題があった為であると仮定する．ここで，Webページから基底の単語を選出し，その有効性を検証する．基底単語選出および実験評価方法としては単語クラスタリングを行う．Webページから得られた単語をクラスタリングし，基底単語を得る．そして新聞記事から得られた基底の単語と，Webから得られた基底の単語とのクラスタリング結果を比較し評価する．また，新聞記事から得られた基底の単語と，新聞記事をコーパスに用いた場合のクラスタリング結果とも比較し，評価を行う．

2 単語クラスタリング

2.1 ベクトル空間モデル

単語クラスタリングは，単語のベクトル空間モデルに基づいて行なわれる．ベクトル空間モデル作成の流れは以下の通りである．

1) 特徴ベクトル作成

2) 属性ベクトル作成

3) 類似度計算

2.1.1 特徴ベクトル作成

単語クラスタリングを行う際に，単語を特徴ベクトル化する必要がある．テキストデータから注目する単語に対応する表層的な特徴を抽出し，個々の特徴の重みを表わす数値を特徴ベクトルの要素とする．クラスタリングの対象である n 語を $w_1, w_2, \cdots, w_n$ で表わし，単語を表現する特徴数を m とした場合，単語 w_i の特徴ベクトルは次式で表せる．

$$w_i = (v_{i1}, v_{i2}, \cdots, v_{im})$$

ここで $v_{ij}(j = 1, 2, \cdots, m)$ は，j 番目の特徴と単語 w_i の関係の強さを表す数値で，特徴の重みと呼ぶ．

2.1.2 属性ベクトル作成

特徴ベクトルはテキストデータから抽出されるが，テキストデータから得られる特徴には常に欠落やノイズが存在するため，適切な判別を行うことができない．そのため，特徴間の関連性を考慮して，主要で相互に関連性の少ない特徴へ変換し，テキストデータ中の特徴のノイズや欠落の影響を減らす必要がある．変換したベクトルを属性ベクトルと呼ぶ．

特徴の線形変換によって，特徴行列 G_0 の m の特徴を，$k(\leq m)$ の互いに関連性の少ない特徴行列，つまり個々の単語の属性の重みを要素とする属性行列 G へ変換できると仮定すると，その変換は m 行 k 列の変換行列を掛け合わせることに等しい．

$$G = G_0 K$$

また，K を特徴ベクトルに掛け合わせることで属性ベクトルに変換できる．属性ベクトルは次式で表わすことができる．

$$w_i^{'} = w_i K = (v_{i1}^{'}, v_{i2}^{'}, \cdots, v_{ik}^{'})$$

2.1.3 類似度計算

類似度とは，2つの単語の似ている度合を表わす尺度であり，値が大きいほどその単語同士が類似していることを表わす．単語 W_p と W_q の類似度 $sim(W_p, W_q)$ は，以下の式で表せる．

$$sim(W_p, W_q) = \frac{w_p^{'} w_q^{'}}{|w_p^{'}||w_{q|}^{'}}$$

2.2 クラスタリング手法

クラスタリング手法は大きく，階層的手法と分割的手法とに分けられる [2].

2.2.1 階層的手法

階層的手法では，対象間の類似度を指標にし，樹状の分類構造を構成することを目標とする手法である．分類構造を適当な箇所で切ることにより，任意の個数のクラスタを得ることができる．樹状の分類構造であるため，切断箇所所が根に近づくほど多数の構成単位のクラスタが含まれる．つまり階層的構造を持っている．今回の実験では，階層的手法として Ward 法を用いる．Ward 法の代表的な距離関数 $D(C_1, C_2)$ を表す．

$$D(C_1, C_2) = E(C_1 \bigcup C_2) - E(C_1) - E(C_2)$$

ただし，

$$E(C_i) = \sum_{x \in C_i} (D(x, c_i))^2$$

2.2.2 分割的手法

分割的手法は，分割の良さの評価関数を定め，その評価関数を最適にする分割を探索する．可能な分割の総数は単語数に対して指数的なので，実際は準最適解を求める．今回の実験では，分割的手法として k-means 法を用いる．代表的な k-means 法では，セントロイドをクラスタの代表点とし，

$$\sum_{i=1}^{k} \sum_{x \in C_i} (D(x, c_i))^2$$

の評価関数を最大化する．

3 基底単語選出

本実験で行った基底単語選出の流れを示す．

step1 Web 文書から本文のみを抽出

step2 本文中から文章のみを抽出．ここでは句読点を含む文を文章と定義する

step3 文章中から単語を抽出

step4 単語間の類似度より類似している単語をまとめる

step5 類似している単語群から最も頻度の高い単語を基底単語として選出

基底単語選出の際，単語間の類似度を測る必要がある．単語 $w1$ と $w2$ の類似度は，$w1$ の頻度を $f1$，$w2$ の頻度を $f2$，$w1$ と $w2$ が共起した頻度を $f3$ とするとき，次式で求められる．

$$sim(w1, w2) = \frac{2 \cdot f3}{f1 + f2}$$

これは Dice 係数と呼ばれている．ここで，頻度は文書数である．つまり，$f1$ は $w1$ を含む文書の数である．また，$f3$ は $w1$ と $w2$ をともに含む文書の数である．

以上より得られた 100 個の基底単語を表 1 に示す．この 100 単語を用いて単語クラスタリングを行う．

表 1: Web 文書から得た基底単語

仮装	選択	登録	現在	医療
写真	東	機能	ァ	状況
トップ	方法	電話	セット	名
系	店	水	機関	経営
先	米	管理	関連	環境
ノ	コンピュータ	海	分	形
地区	参加	事業	所	章
日本	実験	処理	面	発表
案内	担当	指導	私	笑
概要	インターネット	料	時	生産
交流	構造	一覧	研究	教育
関係	手	市	書	問題
学会	クリック	今日	番号	情報
システム	璽	集	大阪	方
質問	法	縺	場合	基礎
紹介	任	量	掲載	靴
計画	ファイル	掲示板	一般	垢
子	目的	下	長	里
株	学生	実現	イ	心
データ	ニ	会	文字	利用

4 実験

ここでは，Web をコーパスに用いた単語クラスタリングを行う．新聞記事から得られた基底の単語と，Web から得られた基底の単語とのクラスタリング結果を比較し評価する．また，新聞記事から得られた基底の単語と，新聞記事をコーパスに用いた場合のクラスタリング結果とも比較し，評価を行う．実験で用いる単語は以下に示す 25 単語である．

動物 プードル，チワワ，犬，猿，ゴリラ

動物 カレー，ラーメン，スパゲッティー，焼きそば，ハンバーグ

感情・人生観 幸福，満足，愛情，結婚，運命

繁栄しているもの 情報，知識，手段，交通，設備

地名 今帰仁，沖縄，ブリスベン，オーストラリア，オーストリア

選出した基底単語 $(v_1, v_2, \cdots, v_{100})$ と対象の単語 w の特徴ベクトルを得るのに，“w v_i” をクエリにして google から検索を行なう．[3] そのヒット数を h_i とおく．今回，共起頻度として検索ヒット数を用いる．そして対象単語 w の特徴ベクトルを以下で表現する．

$$w = (\frac{h_1}{Z}, \frac{h_2}{Z}, \cdots, \frac{h_i}{Z}, \cdots, \frac{h_{100}}{Z})$$

ここで Z は w を正規化する定数である．

$$Z = \sqrt{\sum_{i=1}^{100} h_i^2}$$

以上を用いて単語クラスタリングを行なった．結果を図 1，2 に示す．

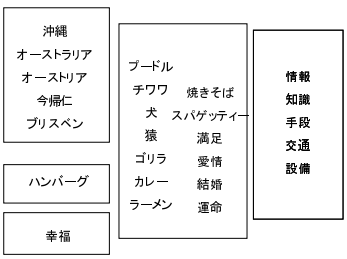

図 1: ward 法 (基底：Web, コーパス:Web)

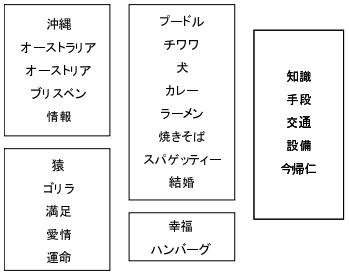

図 2: k-means 法 (基底：Web, コーパス:Web)

また，比較のために新聞記事から得た基底単語と，Web をコーパスに用いたクラスタリング結果を図 3，4 に示す．

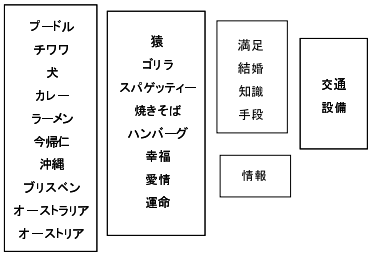

図 3: ward 法 (基底：新聞記事, コーパス:Web)

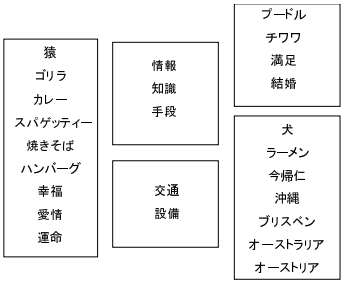

図 4: k-means 法 (基底：新聞記事, コーパス:Web)

Web から得た基底単語と Web をコーパスにした場合のクラスタリング結果と，新聞記事から得た基底単語と Web をコーパスにした場合のクラスタリング結果を比較した場合，Web を基底単語にした場合の方が良い実験結果が得られている．

同様に，新聞記事から得た基底単語と，新聞記事をコーパスに用いた場合のクラスタリング結果を図 5，6 に示す．ここでは毎日新聞記事 1 年分を使用している．

Web から得た基底単語と Web をコーパスにした場合のクラスタリング結果と，新聞記事から得た基底単

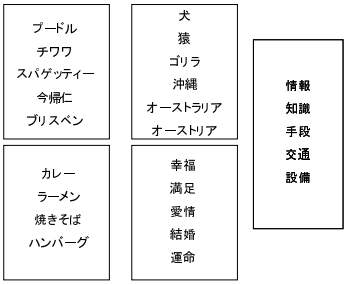

図 5: ward 法 (基底：新聞記事, コーパス:新聞記事)

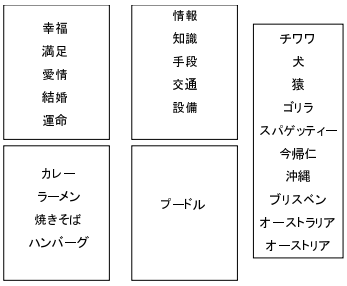

図 6: k-means 法 (基底：新聞記事, コーパス:新聞記事)

語と新聞記事をコーパスにした場合のクラスタリング結果を比較した場合，新聞記事を基底単語にした場合の方が良い実験結果が得られている．

5　考察

新聞記事から得た基底単語と Web をコーパスにした場合のクラスタリング結果をみると，ward 法で行ったクラスタリングでは，適切にクラスタリングされているクラスはなく,「地名」の概念は同じクラスに分類されてはいるが，同じクラス内に「動物」の概念の単語と「食べ物」の概念の単語とが入り混じっている．また，k-means 法で行ったクラスタリング結果を見ても，ward 法と同様に他の概念の単語が入り混じっている．

しかし，Web から得た基底単語と Web をコーパスに用いた場合のクラスタリング結果をみると，ward 法で行ったクラスタリング結果では,「動物」,「食べ物」,「感情・人生観」のクラスの単語が入り混じっているが,「地域」,「繁栄しているもの」のクラスは適切にクラスタリングされている．また，k-means 法で行った場合は，適切なクラスがなく，他の概念の単語が入り混じっている結果になった．よって，新聞記事から基底単語を得るよりも，Web から基底単語を得る方が良いと考えられる．

新聞記事から得た基底単語と新聞記事をコーパスに用いた場合，プードル，チワワ，スパゲッティー，今帰仁，ブリスベンの単語にはスパース性の影響があるため，適切にクラスタリングされないことが報告されている [1]．しかし今回の実験では，ward 法では今帰仁，ブリスベンを含む「地名」の概念が適切にクラスタリングされており，プードル，チワワ，犬という関連性の深い単語が同クラスにクラスタリングされている．k-means 法の場合でも同様に，プードル，チワワ，犬は同クラスにクラスタリングされている．このことから，スパース性の影響は解消されたと考えられる．

しかし，精度としては新聞記事から得た基底単語と新聞記事をコーパスに用いた場合の方が良い．これは，Web から得た基底単語の個数に問題があると考える．今回の実験では，基底単語を 100 単語と設定したが，より最適な単語数，つまりより最適な特徴数を選べば，精度向上に継がると考える．また，今回基底単語を得るために用いた Web データの量にも問題があると考える．今回，Web データの一部から基底単語を得た．データ量を増やし，基底単語を修正することにより，より精度が向上すると考える．

6　おわりに

本研究では Web ページを用いた基底単語選出を行った．

コーパスのスパース性を解消することができ，さらに，新聞記事から得た基底単語よりも良い結果となることが確認できた．しかし，精度としては新聞記事をコーパスにした場合に劣るため，今後の課題としては，さらなる精度向上に努める必要がある．その方法として，最適な特徴数の選出を考える．また，基底単語を得るための Web データを増やす必要がある．

参考文献

[1] 大城亜里沙，新納浩幸，佐々木稔：“検索エンジンを利用した単語クラスタリング”，言語処理学会第 10 回年次大会,pp17-20 , (2004).

[2] 神嶌敏弘：“データマイニング分野のクラスタリング手法 (1) ークラスタリングを使ってみよう!ー”，人工知能学会誌，Vol18,No.1，pp.59-65, (2003).

[3] 藤井丈明，新納浩幸，佐々木稔：“語義識別の誤り原因の調査とオンザフライの類似語判定”，言語処理学会第 10 回年次大会,pp753-756 , (2004).