ＴＦ・ＩＤＦ重み付けを用いた中国語意味マップの自己組織化

馬 青[1] ZHANG Min[2] ZHOU Min[3]
[1]通信総合研究所 [2]Tsinghua University [3]Microsoft Research Asia
qma@crl.go.jp

1 はじめに

単語間の意味的類似度を計算する技術は，情報検索や多義語の曖昧性解消など多くの自然言語処理のタスクに応用できる．これまでコーパスに基づく統計的手法が多数提案されてきた（例えば[1, 2]）．このようなアプローチによる単語の類似度計算は一般的に以下のような手順で行われる．まず，対象単語を共起語のセットで表現する．次に，このような単語表現を何らかの方法でベクトルに変換する．最後に，単語間の類似度をベクトル間の距離で計算する．

情報検索などへの応用には単語間の類似度だけでなく単語のグローバル的なソーティングが必要になる．これについては，従来から種々のクラスタリング技術を用いて行われていたが，クラスタリング手法はその結果に可視性を欠くなどの問題がある[3]．そのために，クラスタリングの代わりに単語を，意味的類似性を距離とする連続した意味空間（つまり，意味的に近い単語どうしは近いところに，意味的に遠い単語どうしは離れたところに配置されるような空間）にマッピングする技術が必要である．このような意味空間にマッピングされた結果を意味マップと呼ぶ．

これまで英語と日本語においては，このようなマップをコーパスに基づくコネクショニストアプローチによって自動構築する研究がいくつかなされてきた[4, 3]．英語の場合，共起語は目標単語の前後にある２単語とし，目標単語のコーディング（つまり，ベクトルへの変換）にはランダムコーディング法を用いている．それに対し，日本語の場合，共起語は目標単語（名詞）と文法関係（名詞句）にある形容詞と形容動詞とし，目標単語のコーディングには単語の類似度を考慮に入れた相関コーディング法を用いた．意味マップの自己組織化を単語の類似度計算に基づく一種のソーティングとして捉えるなら，コーディングにおいて，如何に有効な単語類似度計算を取り入れるかが良い意味マップの自動構築の鍵となるであろう．その意味では日本語に用いられた相関コーディング法は英語に用いられたランダムコーディング法より進歩していると言える．しかし，このコーディング法では共起語の重要度を測る重要なパラメータである共起頻度を取り入れていない．また，これまでの研究では作成した意味マップの評価はすべて数値を用いた客観的なものではなく直観的に行われていた．

本稿は日本語意味マップの自動構築の継続研究として，中国語名詞の意味マップの自己組織化について述べる．中国語の場合，形容詞/名詞－名詞で名詞句を構成しているため，共起語はそのような文法関係にある形容詞と名詞とする．目標単語のコーディングには単語の類似度を組み込んだ相関コーディング法に加え，ＴＦ・ＩＤＦ重み付けを用いた新しいコーディング法を提案する．そして，意味マップの評価は直観によるもの，クラスタリングと多変量解析を用いた比較によるものに加え，新しく導入した精度，再現率，そしてF-measureスコアを用いた数値評価も行う．

2 自己組織化神経回路網モデル

意味マップの自動構築マシンとしてはKohonenの自己組織化神経回路網モデル（Self-organization Map，略してSOM[5]）を用いる．SOMは高次元入力を持つ２次元配列のノードで構成され，以下に述べる自己組織化によって，高次元データをその特徴を反映するよう２次元空間ににマッピングすることができる．

入力$x = [\xi_1, \xi_2, \cdots, \xi_n]^T \in \Re^n$ならば，個々のノード$i$はそれぞれ参照ベクトル$m_i = [\mu_{i1}, \mu_{i2}, \cdots, \mu_{in}]^T \in \Re^n$を持つものとする．但し，参照ベクトルの要素$\mu_{ij}$はノード$i$と入力要素$\xi_j$の間の重みであり，自己組織過程において少しずつ修正される．入力ベクトルxが与えられたとき，まず，その入力をすべてのノードの参照ベクトルと比較し，ユークリッド距離の一番短いノードを活性化する．マッピング処理段階ではこのノードのみ活性化される．このノードを勝者ノードと呼ぶ．即ち，勝者ノードcは以下のように選ばれる．

$$c = argmin_i\{||x - m_i||\} \quad (1)$$

一方，自己組織化過程では，グローバルに自己組織化が行われるように，勝者ノードだけでなくその近傍のノードも活性化させ，リラックス処理を行う．即ち，活性化されたすべてのノードに対し，それらの参照ベクトルを入力ベクトルに近づくように修正を行う．

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \quad (2)$$

ここで，tは学習回数で，$h_{ci}(t)$は例えば以下のように

定義された近傍関数である．

$$h_{ci}(t) = \alpha(t) \cdot exp(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}) \quad (3)$$

但し，$r_c \in \Re^2$ と $r_i \in \Re^2$ はそれぞれ勝者ノード c と近傍ノード i の位置ベクトルである．従って，項 $||r_c - r_i||$ は近傍ノード i が勝者ノード c から離れて行くにつれ，h_{ci} が小さくなり $m_i(t)$ の修正量が小さくなることを意味する．また，$\alpha(t)$ は学習率で，$\sigma(t)$ は近傍の大きさ（半径）である．これらは時間と共に単調に減少していく関数であればよい．

通常，学習過程は「整列」フェーズと「微調整」フェーズからなる.「整列」フェーズにおいては $\alpha(t)$ と $\sigma(t)$ の初期値を共に大きく取り，時間と共に減少して行く．ノードの配置の基本形はこのフェーズで形成される．一方，残りのフェーズでは，$\alpha(t)$ と $\sigma(t)$ は小さい値のまま長時間をかけて，初期フェーズで形成された基本形を微調整する．

3 データコーディング

n 個の単語 w_i $(i = 1, \cdots, n)$ が存在し，それらの意味マップを構築すると仮定する．単語 w_i は以下のように共起語のセットで定義することができる．

$$w_i = \{a_1^{(i)}, a_2^{(i)}, \cdots, a_{\alpha_i}^{(i)}\} \quad (4)$$

但し，$a_j^{(i)}$ は w_i との j 番目の共起語で，α_i は w_i と共起する単語の数である．要素 d_{ij} が単語 w_i と w_j の意味的類似度であるような相関行列 D が存在すると仮定すれば，個々の単語 w_i を相関行列 D の i 行目の要素で構成される多次元ベクトルにコーディングすることが考えられる．すなわち，

$$V(w_i) = [d_{i1}, d_{i2}, \cdots, d_{i\omega}]^T \quad (5)$$

ここで，$V(w_i) \in \Re^n$ はSOMの入力である．即ち，SOMの役割は，このような多次元ベクトルを，自己組織化によって，それらの間に存在する意味関係を顕在化させて，2次元的に表現することである．従って，良いコーディング法を得るには，単語間類似度 d_{ij} の求め方がキーポイントとなる．

3.1 従来法

従来法においては，単語 w_i と w_j の間の類似度 d_{ij} は以下のように計算される．

$$d_{ij} = \begin{cases} \frac{(\alpha_i - c_{ij}) + (\alpha_j - c_{ij})}{\alpha_i + \alpha_j - c_{ij}} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

但し，α_i と α_j はそれぞれ w_i と w_j が共起する単語の数で，c_{ij} は w_i と w_j に共通する共起語の数である．従って，ここでの d_{ij} は単語 w_i と w_j の間にどれだけ共通の共起語を持つかという意味での両単語間の正規化された意味的距離である．d_{ij} が小さければ，両単語は意味的に近く，d_{ij} が大きければ，両単語は意味的に遠い．

3.2 ＴＦ・ＩＤＦ重み付け法

この方法は，特定の目標単語との共起頻度が高く他の目標単語との共起頻度が低い共起語こそ，特定目標単語にとって重要な共起語であるという仮説に基づくものである．ＴＦ・ＩＤＦ計算はこれまで重み付け方法として，主に文書分類や情報検索に用いられる重要キーワードの選定に用いられてきた．本研究では，共起語のセットで表現される目標単語を文書と見なすことによって，この技術を共起語の重要さの重み付けに利用できるようにした．

この方法においては，単語 w_i と w_j の類似度 d_{ij} は以下のように計算される．

$$d_{ij} = \begin{cases} \frac{(T_i - T_{ij}) + (T_j - T_{ij})}{T_i + T_j - T_{ij}} & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

但し，T_i と T_j はそれぞれ w_i と w_j が持つ共起語の数 α_i と α_j[式 (6)] の拡張で，T_{ij} は w_i と w_j の共通する共起語の数 c_{ij}[式 (6)] の拡張である．これらは以下のように求められる．

$$T_i = \sum_{x=1}^{x=\alpha_i} t_x^{(i)} \quad \text{and} \quad T_{ij} = \sum_{x=1}^{x=c_{ij}} t_x^{(ij)} \quad (8)$$

但し $t_x^{(i)}$ は単語 w_i の共起語 $a_x^{(i)}$ $(x = 1, \cdots, \alpha_i)$ のTF·IDF 値で，$t_x^{(ij)}$ は単語 w_i と w_j $(x = 1, \cdots, c_{ij})$ が共通する共起語 $a_x^{(i)}$ $(x = 1, \cdots, c_{ij})$ の TF·IDF 値である．これらはそれぞれ以下のように計算される．

$$t_x^{(i)} = tf(a_x^{(i)}, w_i) \cdot idf(a_x^{(i)}) \quad (9)$$

$$t_x^{(ij)} = tf(a_x^{(i)}, w_i, w_j) \cdot idf(a_x^{(i)}) \quad (10)$$

ここで，$tf(a_x^{(i)}, w_i)$ は共起語 $a_x^{(i)}$ と単語 w_i の共起頻度，$tf(a_x^{(i)}, w_i, w_j)$ は $a_x^{(i)}$ と w_i と w_j の共起頻度，そして，$idf(a_x^{(i)})$ は $a_x^{(i)}$ が全目標単語と共起する頻度の逆数である．すなわち，

$$idf(a_x^{(i)}) = \log \frac{n}{df(a_x^{(i)})} + 1 \quad (11)$$

但し，n は目標単語の総数で，$df(a_x^{(i)})$ は $a_x^{(i)}$ と共起する目標単語の数である．

3.3 他の方法

式 (8) における TF·IDF 値 $t_x^{(i)}$ と $t_x^{(ij)}$ をそれぞれ共起頻度 $f_x^{(i)}$ と $f_x^{(ij)}$ に置き換えれば，頻度重み付け法が簡単に得られる．また，これまで述べたきた方法では，単語の類似度計算を式 (6) と (7) のような簡単な集合演算で行っている．このような集合演算の代わりにベクトル演算や情報理論を用いる方法も考えられる．これらを用いることによって，更に頻度あるいは TF·IDF 値で構成されるベクトル上の cosine measure，そして頻度で構成されるベクトル上のエントロピー計算を用いた三つの方法を考案し，前述した三つの手法に加え計六つの手法による比較実験を行う．

4 実験結果

4.1 データ

実験結果の評価をより客観的かつ容易に行うために，目標単語（計８５個の名詞）は現代中国語意味分類辞書[6]の六つのカテゴリから選んだ．そして，辞書には載っていないが新聞によく出現する中国人の人名を新たなカテゴリーとして加えた．共起語は中国語の名詞句を形成する形容詞と名詞で，１１年分の人民日報から自動的に収集された．共起語ののべ総数は 69,030 で，異なり総数は 22,118 であった．

4.2 SOM

実験には 13 × 13 のの２次元配列の SOM を用いた．入力の次元 n は対象単語の数と同様，85 であった．整列フェーズにおいては，学習総回数 T を 10,000 に，学習率の初期値 $\alpha(0)$ を 0.1 に，そして，近傍の初期半径 $\sigma(0)$ を 13 に設定した．微調整フェーズにおいては，学習総回数 T を 100,000 に，学習率の初期値 $\alpha(0)$ を 0.01 に，そして，近傍の初期半径 $\sigma(0)$ を 7 に設定した．

4.3 評価方法

数値評価

精度（P）と再現率（R）を以下のように定義し意味マップ（の分類結果）の数値評価を行った．

$$P = \frac{\sum_{i=1}^{C} p_i}{C}, \qquad R = \frac{\sum_{i=1}^{C} r_i}{C} \tag{12}$$

但し，C はクラスの総数で，p_i と r_i はそれぞれ一クラス i を見る場合の精度と再現率である．それらは以下のように定義される．

$$p_i = \frac{\text{クラス } i \text{ として正しく分類された単語の数}}{\text{マップ上のクラス } i \text{ の単語総数}} \tag{13}$$

$$r_i = \frac{\text{クラス } i \text{ として正しく分類された単語の数}}{\text{クラス } i \text{ の単語総数}} \tag{14}$$

直観による評価

意味マップのもっとも顕著な特徴は分類結果の可視性及び連続性にあるので，意味マップの評価としては，数値評価だけでは不十分で，直観による判断も行った．

他手法との比較による評価

提案手法の有効性を見るために，分類能力においては階層型クラスタリング手法との比較を行った．そして，提案手法の必要性を見るために，理論上可視化能力を有す主成分分析といった多変量解析手法が本タスクに適用できるかどうかを主成分の寄与率分析及び計算機実験で検証した．

4.4 結果

表 1に各種のデータコーディング法を用いて得られた意味マップ及び階層型クラスタリング結果の比較を F-measure スコアの順で示す．但し，考案した cosine measure コーディングを用いて得られた意味マップは各クラスの単語が混在して意味のある結果が得られなかったため，数値評価を行わなかった．この表からは TF·IDF 重み付けコーディングを用いて得られた意味マップの分類精度及び再現率がほかのどの場合よりも高いことが分かる．そして，意味マップの分類結果とクラスタリング結果を比較した場合，従来のコーディング法と TF·IDF 重み付けコーディング法のいずれを用いた場合でも，意味マップのそれぞれの場合の F-measure スコアがクラスタリングのそれぞれのそれより高いことが分かる．

表 1: 各種コーディング法及びクラスタリングの比較結果

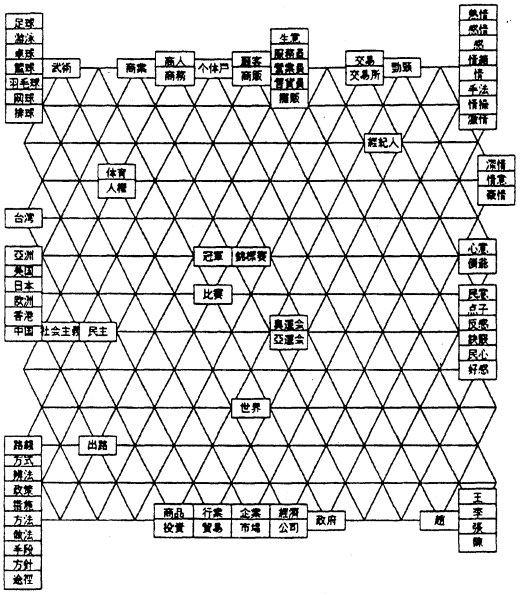
	精度	再現率	F-measure
クラスタリング１ *	0.936	0.864	0.899
エントロピー	0.925	0.874	0.899
従来法	0.926	0.90	0.913
頻度重み付け法	0.928	0.90	0.914
クラスタリング２ **	0.95	0.896	0.922
TF·IDF 重み付け	0.944	0.907	0.925

* 従来のコーディング法を採用した場合
**TF·IDF 重み付け法を採用した場合

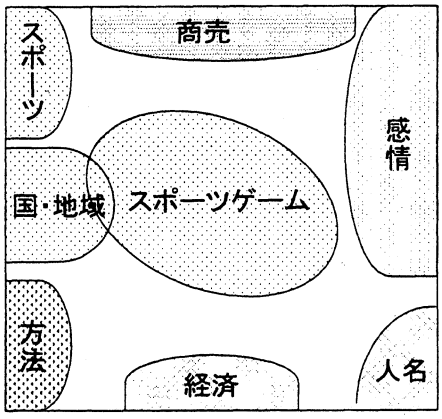
図 1に (a)TF·IDF 重み付けコーディングを用いて得られた意味マップ図 1；(b) これを八つのグループに分けた図を示す．全８５個の名詞のうち，僅か６個の名詞だけが辞書の定義と直観のいずれとも異なるクラスにマッピングされた．しかも，これらの中でも，名詞「勁頭 (闘志)」は正しい領域<感情>に近いところにマッピングされている．そして，<スポーツゲーム>にマッピングされている名詞「世界」は辞書の定義とは異なるものの，直観的には合うところがある．残りの７８の名詞が全部正しく八つのグループに配置されている．但し，すべての名詞はもともと七つのカテゴリから収集されたものである．その違いは，<政策>と<方法>が統合され，<スポーツ>が<スポーツ>と<スポーツゲーム>，そして<商売>が<商売>と経済>に分割された結果である．このような分類結果は明らかに元の分類と矛盾しない．従って，このように自己組織化された意味マップは基本的に中国語辞書と合致していると言えよう．もちろん，これは我々の直観とも合致していると思われる．

意味マップの結果と階層型クラスタリングの結果を比較すれば，クラスタリングの結果においてもすべての名詞が意味マップとまったく同じ八つのグループに分類されており，両者の分類結果が非常に似ていることが分かる．

同じ TF·IDF 重み付けコーディングを用いた場合のデータに対する主成分分析を行った結果，２個の最大の主成分の累積寄与率がわずか 8.29%で，１０個の最大の主成分の累積寄与率もわずか 24.53%であった．一般的には主成分の累積寄与率が 70 ないし 80%以上大きくならないと，多変量データを正しく縮約できないとされている．従って，意味マップの構築に多変量解析を用いることは困難である．実際，データを第１主成分と第２主



(a)



(b)

図 1: 自己組織化された意味マップ．このマップでは名詞「情感 (emotion)」も＜感情＞領域に入っているが，他の名詞に隠されている．

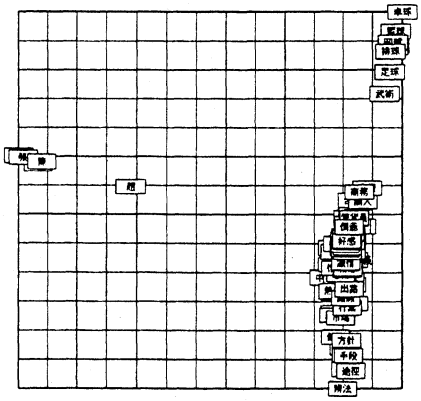


図 2: 主成分分析によって得られた意味マップ

成分を軸とする 2 次元平面にプロットする計算機実験も行ってみた．その結果は図 2に示している．ほとんどの名詞が右側の一部の領域に集中してうまく分類できなかった．

5 結び

本稿は中国語意味マップの自己組織化について述べた．単語のコーディングにはＴＦ・ＩＤＦ重み付けを導入した新しいコーディング法を提案した．意味マップの評価には精度，再現率，そして F-measure スコアといった客観的な数値評価を導入した．新しいコーディング法の有効性は従来の相関コーディング法も含め五つのコーディング法との比較によって確認された．提案した自己組織化手法の有効性及び必要性は従来のクラスタリング手法及び多変量解析手法との比較を通じて明らかにされた．

参考文献

[1] Hindle, D.: Noun classification from predicate argument structures, *ACL'90*, Pittsburgh, PA, pp. 268-275, 1990.

[2] 森 信介，西村 雅史，伊東 伸泰：クラスに基づく言語モデルのための単語クラスタリング，情報処理学会論文誌，38(11)，pp. 2200-2208，1997.

[3] 馬青，神崎享子，村田真樹，内元清貴，井佐原均：日本語名詞の意味マップの自己組織化，情報処理学会論文誌，42(10)，pp. 2379-2391，2001.

[4] Ritter, H. and Kohonen, T.: Self-organizing semantic maps, *Biological Cybernetics*, 61, pp. 241-254, 1989.

[5] Kohonen, T.: Self-organizing maps, Springer, 2nd Edition, 1997.

[6] Dong, D. N., et al. (Eds.), Contemporary Chinese Classified Dictionary, *Han-Yu-Da-Ci-Dian Press*, 1998.