共起関係に基づく階層型単語概念体系の動的構築法

新井 志勇人 森田 和宏 泓田 正雄 青江 順一
徳島大学大学院 先端技術科学教育部

1. はじめに

自然言語処理における重要な知識源の一つに，単語間の上位下位関係や同義関係を記述したシソーラスがあり，意味解析やデータマイニングなどに有用である．しかし，シソーラスの作成・維持および優先語(以後，ディスクリプタと呼ぶ)の付与には，労力がかかるという問題点がある．この問題を解決するために，シソーラスの自動構築，ディスクリプタの自動付与の 2 つをおこなう必要がある[1]．本稿では，シソーラスの自動構築を目的とする．

シソーラスを自動で構築するクラスタリング手法が提案されている[2]．クラスタリングは一般に，共起情報などを属性ベクトルとしたデータ集合に対し，一括で処理をおこなう．そのため，シソーラスに登録されていない単語(以下，未知語と呼ぶ)の追加といった微小なデータの変化に対してコストが高い．また，クラスタ間の距離的関係は存在するが，上位下位関係を考慮しているものは少ない．

そこで，本稿では共起関係を用いて，データの変化にインクリメンタルに対応する階層型の単語概念体系構築法を提案する．

2. クラスタリング手法

クラスタリング手法は大きく，階層的手法と分割最適化手法とに分けられる[3]．階層的手法では，クラスタ間の距離関数に基づき，最も距離の近い二つのクラスタを逐次的に併合し，樹上の分類構造を構成する．分類構造を切ることで，任意の個数のクラスタを得ることができる．分割最適化手法では，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に対して指数的なので，実際は準最適解を求める．

代表的な手法としてそれぞれ，Ward 法，k-means 法が挙げられるが，これらの距離尺度や類似尺度は上位・下位を考慮していないため，自動で階層型のシソーラスを構築するのは困難である．また，未知語の登録に対しても[4]などの研究がなされているが，体系内でもっとも類似しているクラスタに単語を登録するのみである．本システムは，体系内にあまり類似性がない未知語には新しいクラスタを生成する点で[4]とは異なる．

3. 提案手法

3.1 システムの概要

本システムは，共起情報登録部，単語情報（ベクトル）作成部，体系構築部より構成される．

共起情報登録部では，入力として与えられた共起情報から共起辞書を作成する．次にこの共起辞書を用いてベクトル形式の単語情報(以後，単語ベクトルと呼ぶ)を単語情報作成部で作成する．体系構築部では，単語ベクトルを用いて各階層との類似度計算をおこなう．階層内には単語集合を格納するノードが複数存在し，最高の類似度を持つノードに単語を登録し，もしくはノードへ追加することで体系を構築する．各処理の詳細については，次節以降で述べる．

本システムでは，前述の処理を一つのサイクルとし，入力された共起情報ごとに繰り返す．このため，入力のどの時点においても体系が形成されており，最初から構築しなおす必要はなく，順次追加するだけでよい．

3.2 共起辞書

共起辞書は，単語，共起語，共起頻度の 3 つから構成された共起情報で構成する．例えば，単語の“スポーツ”であれば，共起語として“が楽しめ”，共起頻度して“6”といった具合である．なお，本稿で使用する共起情報は，Google N グラム[5]から抽出した．

3.3 ベクトルの作成

単語ベクトルは，共起語と共起頻度から構成される．共起頻度を p_i，共起語を V_i とすると，単語ベクトル W_V は以下の式で表される．

$$W_V = \sum_{i=1}^{n} p_i V_i$$

次に，グループベクトルについて述べる．グループベクトルは，ノードに登録されている単語集合の共起情報を合わせたもので構成される．単語集合に対する共起語の共起頻度の総和を q_i とすると，グループベクトル G_V は以下の式で表される．

$$G_V = \sum_{i=1}^{n} q_i V_i$$

3.4 類似度

類似度計算は，単語ベクトルとグループベクトルを用いておこなう．単語間の上位・下位関係を抽出する技術としてカルバックライブラー距離(以下，KL 距離)[6]を用いる．これは KL 距離が他の類似尺度と異なり対称性をもたないためである．単語ベクトル W_V からの KL 距離 D_{KL} は以下の式で表される．

$$D_{KL}(W_V, G_V) = \sum_{i=1}^{n} W_{Vi} \log \frac{W_{Vi}}{G_{Vi}}$$

また，単語のノードへの追加，吸収処理にはコサイン距離を用いる．追加処理とは，新しいノードを作成し，単語を登録することであり，吸収処理とは，一番類似度が高いノード(以後，最高類似ノード)へ単語を登録することである．コサイン距離 D_C は以下の式で表される．

$$D_C(W_V, G_V) = \frac{\sum_{i=1}^{n} p_i q_i}{|W_V||G_V|}$$

A Dynamic Classification for hierarchical word concepts based on co-occurrence relations
Shuto Arai, Kazuhiro Morita, Masao Fuketa and Jun-ichi Aoe
Graduate School of Advanced Technology and Science, The University of Tokushima

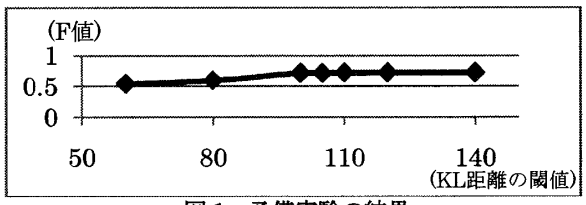

図 1：予備実験の結果

3.5 体系化

本節では，前節までに述べたベクトル，類似度計算を用いた単語の体系化の流れを説明する．なお，体系の初期は階層構造を作る起点(以後，root)のみで構成されている．以下，体系化アルゴリズムを示す．

Step1：単語の除去

登録する単語 a が，体系内に存在すれば一度体系から削除する．a の属しているノード n のグループベクトル G_{Vn} から a の単語ベクトル W_{Va} を取り除き，グループベクトルを $G_{Vn}'=G_{Vn}-W_{Va}$ とする．

Step2：体系の探索

KL 距離 D_{KL} を用いて，単語ベクトル W_{Va} と各ノードのグループベクトル G_{Vi} から類似度と類似方向を求める．$D_{KL}(W_{Va},G_{Vi}) \geq D_{KL}(G_{Vi},W_{Va})$ なら上方向の類似度 $D_{KL}(W_{Va},G_{Vi})$ を取得し，$D_{KL}(W_{Va},G_{Vi}) < D_{KL}(G_{Vi},W_{Va})$ なら下方向の類似度 $D_{KL}(G_{Vi},W_{Va})$ を取得する．

次に，探索方法について説明する．root の下位ノードに対して D_{KL} を求め，一番類似度の高いノードを決定する．決定したノードの下位ノードに対しても同様におこない，最高類似ノード m が決定するまで探索を繰り返す．また，KL 距離の閾値を θ_1 とし，ノード m が $D_{KL}<\theta_1$ なら類似性はないとみなし，step3 で追加処理をおこなう．

Step3：ノードの追加，吸収処理

コサイン距離 D_C を用いて，単語 a に対してノードへの追加処理，吸収処理のどちらをおこなうか決定する．

コサイン距離の閾値を θ_2 とし，$D_C(W_{Va},G_{Vm}) < \theta_2$ なら追加処理をおこなう．追加処理は，Step2 で取得した類似方向が上方向であれば，ノード m と上位ノードの間に新しくノードを作成し，下方向であればノード m の下に新しくノードを作成する．追加したノードに単語 a を登録し，グループベクトル $G_{Va}=W_{Va}$ とする．

また，$D_C(W_{Va},G_{Vm}) \geq \theta_2$ なら吸収処理をおこなう．吸収処理は，ノード m に単語を登録すると同時に，グループベクトル G_{Vm} を $G_{Vm}' = G_{Vm}+W_{Va}$ と更新する．

4. 予備実験

予備実験として，類似度に用いる KL 距離とコサイン関数の閾値を決定する．体系の評価をおこなうために，EDR 電子化辞書[7]から概念を 6 種類選択し，各概念に属する単語が正しく構築できるか確認する．Google N グラムから各概念に属する単語 108 単語(143,819relation)の共起情報を抽出して実験をおこなった．構築された体系に対して，root の各下位ノードから始まる全ての下位ノードをそれぞれクラスタとして再現率，適合率を求め F 値を算出した．再現率・適合率は以下のとおりである．

$$\text{再現率} = \frac{\text{クラスタ内の概念に属する単語数}}{\text{概念に属する単語数}} \times 100$$

$$\text{適合率} = \frac{\text{クラスタ内の概念に属する単語数}}{\text{クラスタ内の単語数}} \times 100$$

表 1：比較実験

	再現率(%)	適合率(%)	F 値
提案手法	50.75	77.88	0.615
k-means 法	73.27	51.26	0.603

図 1 に予備実験の結果を示す．図 1 において，KL 距離の閾値が 110 以上の際に F 値が停滞したため，停滞し始めた値 110 を閾値に決定した．次に，コサイン距離の閾値を変更したとき，閾値 0.36 で F 値が最大となった．この結果を用いて従来手法との評価実験をおこなう．

5. 評価実験

5.1 実験設定

評価実験として，k-means 法との比較をおこない，提案手法の有効性を確認する．概念を 19 種類，各概念に属する単語 333 単語(278,788relation)の共起情報を Google N グラムから抽出して使用する．また，k-means 法は概念名となる語を代表点に選択し，類似度計算にはコサイン距離を用いて実験をおこなった．なお，提案手法，k-means 法の再現率，適合率は 4 節と同様である．

5.2 評価，考察

表 1 に評価実験の結果を示す．F 値は同等の精度となったが，シソーラスの構築としては，適合率が高いほうが良いと考えられる．また，再現率が低い点について考察をおこなう．k-means 法はクラスタ数を固定しているため，類似度が低くてもいずれかのクラスタに分類される．しかし，提案手法では，KL 距離の閾値を満たしていない場合，新しいノードを作成するため，類似性の高いノードが複数存在することがあり，root からの下位ノード数が増加するためである．改善策として，構築した体系のノード同士の距離を測り，距離が近い場合はノードを結合することが挙げられる．

6. まとめ

本稿では，共起関係に基づいて階層型の単語概念体系を動的に構築するシステムの概要について述べた．今後は，体系化したノードの距離関係も考慮するアルゴリズムを考案し，大容量の共起情報を利用して大規模な体系構築をおこなう．また，ディスクリプタの自動付与，単語の曖昧性解消に対してのアルゴリズムを考案することも今後の課題である．

参考文献

[1] 岸田和明：インターネット時代における統制語彙の意義と役割(<特集>統制語彙・シソーラスの現在)，情報の科学と技術，Vol57，No.2，pp.62-67，(2007)

[2] 有田一平，菊池英明，白井克彦：検索語の共起情報を利用した単語クラスタリングと Web 検索への応用，情報処理学会研究報告，Vol.2007，No.76，pp115-120，(2007)

[3] 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) —クラスタリングを使ってみよう!—”，人工知能学会誌，Vol18，No.1，pp.59-65，(2003)

[4] 浦本直彦：コーパスに基づくシソーラス—統計情報を用いた既存のシソーラスへの未知語の配置，情報処理学会論文誌，Vol37，No.12，pp.2182-2189，(1996)

[5] 工藤拓，賀沢秀人，「Web 日本語 Nグラム第1版」，言語資源協会発行(2007)

[6] 別所克人，内山俊郎，片岡良治：単語・意味属性間共起に基づく単語間の階層関係の抽出，電子情報通信学会技術研究報告，Vol106，No.518，pp.31-36，(2007)

[7] 日本電子化辞書研究所 EDR電子化辞書(1995)