1C5-03

シソーラスによる多次元空間への単語配置の補正

Thesaurus-based Adjustment of Arranging Words in a Multi-dimensional Space

笠原 要[*1]
Kaname Kasahara

[*1]日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation

A method of arranging words into a multi-dimensional space using a dictionary and a thesaurus is proposed. In the method, entry words in the dictionary are roughly arranged from frequencies of words in their difinitions. Next, the locations of the entry words in a category of the thesaurus are re-arranged so that they become closer to each other. The method was applied to synonym generation and the result was that the proposed method is better than the original one when recall of the acquired synonyms is high.

1. はじめに

テキストデータを用いて単語を多次元のベクトル形式で表現し，単語の類似性を判別するモデルは“単語のベクトル空間モデル”と呼ばれる (例えば [Hindle 90, Schutze 92, 笠原 97, 稲子 00])．このモデルを利用すると，文や文書等の単語からなる様々な情報をベクトル表現に変換し，ベクトル間の類似度計算を通して一元的に類似性を判別することができる．そのため，人間の類似性判別の工学的な代替として，情報検索 [Schutze 95, 熊本 99]，ナレッジマネージメント支援 [加藤 00]，テキストセグメンテーション [別所 01]，テキストが付与されたマルチメディア情報の可視化 [宮原 02] 等，単語の意味を考慮した知的なテキスト処理に幅広く応用されている．

ベクトル空間モデルでは，利用するテキストデータの種類や規模，作成方法によって，単語に対して作成されたベクトルや単語間の類似度の値が変化するので，モデルの最適化が必要である．しかし，計算される類似度自体が適切であるかどうかは，類似度を利用するタスクによって異なるため，タスクごとにモデルを最適化することが必要である．[笠原 03] では，人間が行う類義語の作成の模擬を基本的タスクと想定し，それに有効なテキストデータ，単語のベクトル空間の作成方法が検討されている．

被験者を用いた心理実験を通して獲得された 200 語の刺激語に対する類義語を評価基準として，作成された類義語と比較することでベクトル空間や作成方式の改善が行われている．その結果を図 1 に挙げる．図中の“国語辞典”は，類義語作成タスクにおいて最適化されたモデルによる様々な再現率における作成精度を表している．テキストデータとしては学研国語大辞典 [金田一 88] を用い，見出し語に対する語義文中の単語の出現頻度に基づいた空間の作成方式 [笠原 97] による単語のベクトル空間を用いている．また，作成される類義語を心理的に馴染みのある単語に限る工夫もされている．その結果，コーパスやシソーラスを用いる従来の作成方式に比べて，どの再現率においても高い精度になっている．図中の“コーパス”は，CD 毎日新聞 2000 の記事をテキストデータとして，同一文内での単語共起の行列を元にして特異値分解を行いベクトル空間を作成する方式 [Schutze 92] による結果である．“シソーラス”は，ベクトル空間モデルではないが，シソーラス（類義語辞典）を用いて同じ分類中の単語同士は互いに類義であると見なす，一般的なモデルによる類義語作成結果を評価したものである．シソーラスとしては，30 万語を 3 千分類した日本語語彙大系 [池原 97] が用いられている．

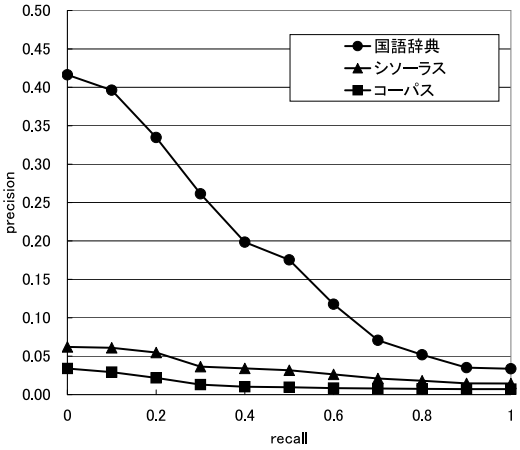

図 1: 再現率-精度曲線

しかしこの評価結果は同時に，最適化された単語のベクトル空間モデルでも，人間の類義語回答の工学的な代替としてはまだ不十分であることを示している．再現率が 0 の時の精度とは，1 語を類義語としての作成した時の正解率にほぼ相当するが，結果としてその値は半分以下 (約 0.4) にとどまっている．これは作成された 1 語の類義語が正しい確率が 40%程度しかないことを示している．また，刺激語に対する基準の類義語の数は約 6 であったので，その語数を作成した場合の精度は評価の 1 つの尺度となる．[笠原 03] の結果は平均 15%の精度にとどまり，6 語中に基準の類義語が 1 語がかろうじて含まれる程度であるため，モデルのさらなる改良が必要であることがわかる．

具体的な改良の方法としては，単語のベクトル空間の作成方式やそれを類義語作成タスクに適応する方式の改良など様々考えられるが，その中でも重要なのは，異なる複数のテキストデータを相補的に利用することと考えられる．それぞれのテキストデータの種類の特質を考慮することができれば，単純な加算的利用以上の効果が期待される．そこで本稿では，[笠原 03] で最適化された，国語辞典に基づくベクトル空間モデルにシソーラスを利用して単語の空間配置を補正する方法を提案する．

連絡先: 笠原 要　日本電信電話株式会社, NTT コミュニケーション科学基礎研究所 〒 619-0237 京都府相楽郡精華町光台 2-4
email:kaname@cslab.kecl.ntt.co.jp

2. ベクトル空間中の単語の再配置

まず，国語辞典からベクトル空間を作成する [笠原 97] の方式および，その結果を用いて類義語を自動作成する [笠原 03] の方式について説明する．次に，ベクトル空間中の単語の配置をシソーラスを用いて補正する試みについて説明する．

2.1 国語辞典からの単語のベクトル空間の作成

ベクトル空間中では個々の単語は，特徴の重み (≥ 0) を要素とするベクトル（「特徴ベクトル」）で表現される．辞典に含まれる個々の見出しの単語 W_i $(i = 1, \cdots, n)$ の特徴ベクトル $Word_i$ は以下の通りとなる．

$$Word_i = (v_{i1}, v_{i2}, \cdots, v_{in}) \tag{1}$$

特徴として，全ての見出し語 (n 語) を用いており，全ての特徴ベクトルは，特徴の重みを要素とする n 行 n 列の行列 (G_1) となる．そして，n 語の特徴を m 個のカテゴリーに分類するシソーラス (類語辞典) を用い，同じ分類に含まれる特徴をカテゴリーに変換して，同じ特徴の重みを同士のみを比較すれば類似度計算できるようにする．

$$Word'_i = (v'_{i1}, \cdots, v'_{ik}, \cdots, v'_{im}) \qquad v'_{ik} = \sum_{l=1}^{n} v_{il} T(l,k) \tag{2}$$

$T(l,k)$ はシソーラスを表す関数で，l 番目の属性が k 番目のカテゴリーに含まれる時は 1，それ以外は 0 を取る．(l,k) の要素の値を $T(l,k)$ とした n 行 m 列の行列を T とすれば，シソーラスで特徴を一般化した特徴行列（「属性行列」と呼ぶ）全体は，n 行 m 列の行列 ($G_2 = G_1 T$) となる．これを属性行列と呼ぶ．

特徴の重みは，国語辞典の見出し語に対する説明文中の単語の出現頻度に基づいて獲得する．獲得方式の詳細は，文献 [笠原 97] を参照されたい．なお，獲得された属性ベクトルそれぞれについて，個々の重みは正規化しておく ($\sum_{k=1}^{m} v'^2_{ik} = 1$)．これを用いて，ベクトル空間に配置された単語 W_i，W_j $(1 \leq i, j \leq n)$ の類似度 sim $(0 \leq sim \leq 1)$ を対応する属性ベクトル $Word'_i$，$Word'_j$ のなす角度の余弦で表す．

$$sim(W_i, W_j) = Word'_i \cdot Word'_j = \sum_{k=1}^{m} v'_{ik} v'_{jk} \tag{3}$$

本稿では，学研 国語大辞典 [金田一 88] と 30 万語を 3000 カテゴリに分類したシソーラス [池原 97] を用い，辞典の約 88,633 語の見出し語が 3000 次元のベクトル空間に配置されたデータ [永森 00] を利用する．

単語のベクトル空間を用いた類義語作成としては，刺激語の属性ベクトルに対して対象とする語彙中の単語それぞれの属性ベクトルとの類似度を計算し，その値の高い単語を類義語として出力することが一般的である．[笠原 03] ではさらに，対象とする語彙として国語辞典の見出し単語を全て用いるよりも，なじみのある単語に限定することで，より精度高く類義語を作成できるという特性を明らかにした．具体的には，心理実験によって定められた単語のなじみの程度を表す主観的特性値である単語親密度が利用されている．刺激語が被験者にとってどの程度なじみがあると感じられるかを表した尺度であり，1 から 7 までの値をとる (1:なじみがない, 7:なじみがある)．約 8 万の語彙について，40 名の被験者を用いた心理実験を通して得られた単語親密度が付与されているデータベース（日本語の語彙特性 [天野 00b]）が存在している．これを用いて 94%以上の成人が知っていると推測される単語親密度 5 以上の値を持ち，ベクトル空間に含まれた，26,371 単語のみから類義語を選択した場合，全ての語彙を用いた場合に比べて作成精度が 1.5 倍以上向上することが報告されている [笠原 03]．

2.2 単語のベクトル空間の補正方式

上記の国語辞典を用いた単語のベクトル空間は，単純にシソーラスの分類を利用する方法やコーパスによる単語のベクトル空間に比べて類義語の作成精度が格段に高いが，はじめに述べた通り，正解となる基準の類義語の多くが作成結果に含まれていない．その原因は，一部の単語の空間中の配置が適当ではないためと考えられる．[笠原 97] では，単語の特徴ベクトルを作成する際に，辞典中の見出し語と，語義文中の単語や関係する見出し語の関係等が考慮されている．しかしそれでも語義文を単語の集まり (“a bag of words”) として扱っているため，idf を用いても語義文特有の言い回しや直接関係無い語義文に含まれるノイズとなる単語を全て除去することはできない．さらに辞典は，人間の利用を想定しているため数語のみの簡潔な記述の語義文も多くあるため，一部の単語のベクトルは正しく配置されていない恐れがある．類義語作成の精度が十分高くないのは上記の理由なのか，あるいは，ベクトル空間のモデルの限界であるかについては，基準の類義語を正解とした分析を行い，それに基づいて配置の補正することを試みることも考えられる．しかしし，ベクトル空間に配置された全ての単語についての類義語を被験者実験から収集することは，時間とコストの面から見て実際的には困難である．そこで本稿は，概念のカテゴリー化に関して心理学で提唱されているプロトタイプ理論 [Rosch 75] の考え方を利用した単語配置の修正の可能性について検討する．

ヒトの記憶の意味カテゴリーの構造として，カテゴリー（分類）を構成する複数の概念（単語の意味）に多く共通する典型的な特徴から構成されるプロトタイプの概念が存在し，カテゴリーに含まれる概念は，これに対する特徴の共通の度合いに基づいて配置されているというモデルが心理実験の結果から提案されている [Rosch 75]．本稿のベクトル空間モデルでは，単語を特徴の重みを要素としたベクトルで表現して互いに類似している単語ほど近くに配置するので，理想的にはこの理論にそった単語配置がされていると期待される．もしも各分類のプロトタイプを自動獲得することができるならば，メンバとなる単語をそのプロトタイプの近辺に再配置する補正が可能となる．しかし，プロタタイプは心理実験に基づいて提案された概念であるため，それを工学的に再現することが必要となる．そこで既存のシソーラスを用いて，その分類に含まれる単語の配置の重心としてプロタタイプを表すベクトル表現することを試みる．勿論，一部の単語の配置はそもそも正しくはないために，作成されるプロトタイプのベクトルにもその影響があるが，カテゴリーに含まれる単語の配置の大多数が適当であるならば，その影響は少ないと予想されるので単純にプロトタイプを表現する．

そのために，ベクトル空間に配置された n の単語全て，あるいはその一部を分類したシソーラス $T2$ を用いる．

$$\begin{aligned} T2 = \quad & \{c_1, \cdots, c_q, \cdots, c_u\} \\ & c_q = \{W_{q1}, \cdots, W_{qr}, \cdots, W_{qs}\} \qquad s \leq n \end{aligned}$$

c_q はシソーラス $T2$ 中の分類の 1 つであり，それに含まれる個々の単語は n 語の単語のいずれかである．$T2$ は，単語の特徴ベクトルを属性ベクトルに変換する際に用いられたシソーラス T とは役割が異なるため，同一のものをを用いてもその影

響は無いと考えられる．シソーラスの分類を構成する単語の重心で分類 c_q のプロトタイプを表すベクトル PT_{c_q} を定義する．

$$PT_{c_q} = \frac{\Sigma_{l=1}^{q} Word'_{cl}}{\|\Sigma_{l=1}^{q} Word'_{cl}\|} \quad (4)$$

このプロトタイプにメンバの単語の配置が近くなるように補正を行えば，それら単語は，分類に含まれない単語よりも互いに近づくことが期待され，同時に，カテゴリーメンバで適切に配置された単語間の相対的な配置は大きく変化しないと考えられる．補正の方法としては様々考えられるが，まず，個々のメンバとプロトタイプの重心を補正した配置と考える (図 2)．分類 c_q に含まれる単語 W_{qr} の属性ベクトル $Word'_{qr}$ の補正 $Word''_{qr}$ は以下で表される．

$$Word''_{qr} = \frac{Word'_{qr} + PT_{c_q}}{\|Word'_{qr} + PT_{c_q}\|} \quad (5)$$

複数の分類に含まれる単語については，個々の分類ごとに複数の補正された単語のベクトルを作成する方法が考えられる．しかし，その場合，1 つの単語の複数のベクトルの曖昧性を解消して選択する方法を検討する必要があるために，まずは単純に 1 単語が 1 配置となるような方法を検討する．すなわち，含まれるカテゴリーが複数ある単語の場合は，まずそれらプロトタイプのベクトルの重心を求め，その後で単語の属性ベクトルと平均した結果を補正した単語の配置とする．

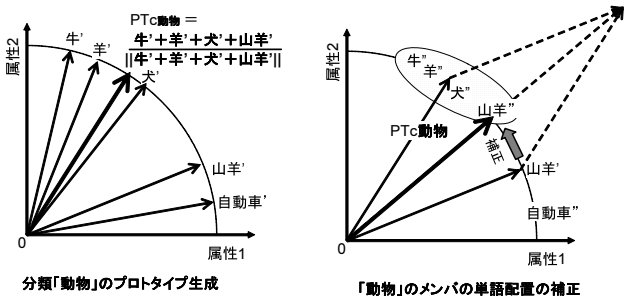

図 2: 提案手法

3. 実験

シソーラスを用いた単語のベクトル空間中の単語の補正が有効であるかを評価するための実験を行った．[笠原 03] で用いられた 26,371 語の単語のベクトル空間 (約 2715 次元) を用いた．シソーラスとしては，特徴ベクトルから属性ベクトルに変換する際に用いた [池原 97] を再度用い，(5) 式により属性ベクトルを補正した．補正したベクトル空間を類義語作成で評価した．刺激語に対して 26,371 語との類似度を計算し，その類似度の高い順に類義語として出力した．評価基準の類義語としては，[笠原 03] で用いられた 200 語の刺激語に対する類義語を用いた．これは，各刺激語に対して，100 人の被験者の中で 2 名以上が列挙した類義語候補について，別の 76 名の被験者の半数以上が妥当と判定した語から成り，1 刺激語あたり平均 6 語の類義語が含まれている．

評価結果を表 1 および図 3 に示す．P_{6w} とは，類義語を 6 語出力したときに，その中に基準の類義語が含まれる割合であり，類義語作成タスク自体を評価する評価値である．P_{avg} は，出力した類義語と基準の比較において，0 から 1 までの 0.1 きざみの再現率における作成精度を計算し (trec eval[Buckley 92])，それを平均化した値であり，6 語以上の類義語を出力させる場合も考慮した，モデルの全般的なパフォーマンスを評価するための値であり，どちらの評価値も 0 から 1 の値を取り得る．P_{6w} は，シソーラスによる配置の補正により，値が下がってしまっているが，P_{avg} では，1 割程度の向上が見られる．これは，' 国語辞典+シソーラス' の再現率-精度曲線を見てわかる通り，再現率が低い時には補正しない場合 (' 国語辞典') よりも精度が若干低くなっているが，再現率が高いときには一様に精度が向上していることで説明できる．

表 1: 類義語作成精度

テキストデータ	P_{6w}	P_{avg}
国語辞典+シソーラス	0.142	0.232
国語辞典	0.145	0.190
テキストコーパス	0.011	0.028
シソーラス	0.006	0.034

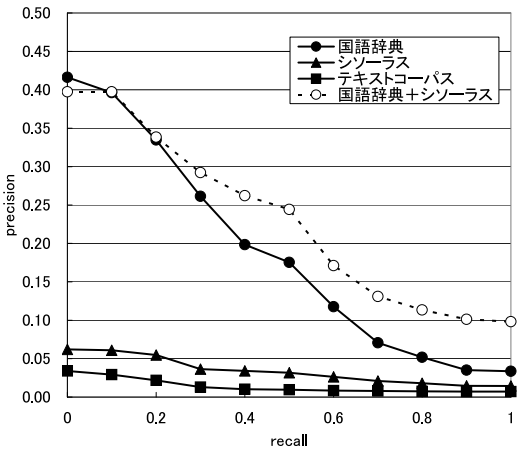

図 3: 単語の再配置による類義語作成の再現率-精度曲線

補正によって再現率が低いときの精度がむしろ低下した理由は，多義の単語の表現が不十分であったためである．表 2 は刺激語「街」に対して作成された類義語である．基準の類義語は「都市」,「市街」,「都会」,「タウン」の 5 語であった．補正しないベクトル空間モデルでは，この内 3 語の正解が表中に含まれているが，提案手法 (“国語辞典+シソーラス”) では，正解の「市街」が含まれていない.「街」はシソーラス [池原 97] では,「都市」と「道路」の分類に含まれる一方,「市街」は「用地」,「道路」,「都市」の分類に含まれており，一致しない分類がある．そのため補正によって,「街」と「市街」の類似度が相対的に低下したためと考えられる．この問題を解決するためには，それぞれの分類を単純に考慮して作成した補正された 2 つの単語のベクトルを作成し，それを別の概念と見なすことが考えられる．そして，類義語作成を行う場合には，一方が作成リストに現れた場合に，もう一方は重複して出力しないような仕組みを行えば，この問題が解消されると予想される．

本稿では，プロトタイプのベクトルの作成の方法および，それを用いた単語の属性ベクトルの補正として非常に単純に行ったため，補正には改善の余地が多くある．多義の単語の配置の補正については上記に述べた通りであるので，ここでは，プロ

表 2: 類義語作成例 (刺激語:「街」, 語彙数=26,371)

	国語辞典+シソーラス		国語辞典	
順位	検索語	類似度	検索語	類似度
1	都市	0.892	市街	0.665
2	タウン	0.883	都市	0.637
3	ベッドタウン	0.8	ベッドタウン	0.596
4	田園都市	0.876	タウン	0.595
5	近郊	0.866	田園都市	0.578
6	繁華街	0.848	近郊	0.557
7	国際都市	0.845	大通り	0.556
8	古都	0.841	東海道	0.527
9	京	0.837	市	0.521
10	町	0.834	繁華街	0.520

トタイプの作成について考察する.

実験では, プロトタイプを分類に含まれる全ての単語の属性ベクトルの平均と考えたが, その各ベクトルの補正を行うことを主眼としているため, 配置があまり正しくない単語, すなわち, 極端に分類中の個々の単語から離れている単語は, プロトタイプのベクトルの計算に用いるべきではないと考えられる. また, 全ての属性をそのまま平均化するよりも, 相対的に値が大きな属性のみを考慮し, それ以外の属性の重みを 0 としてプロトタイプを作成した方が, プロトタイプ理論の考え方に沿っていると考えられる. これらのモデルの様々なパラーメターを考慮することにより, 今回の実験により, ベクトル空間モデルの可能性を明らかにすることができたと考えられる.

4. おわりに

本稿では, テキストデータを用いて単語を多次元空間に配置するベクトル空間モデルについて, その配置をシソーラスを用いて補正することが有効であるか検討した. シソーラスの分類に含まれる単語のベクトルよりその分類のプロトタイプを表すベクトルを作成し, そのベクトルに分類のメンバが近づくような単純な方式で配置を補正し, 類義語作成のタスクにおいて評価を行った. その結果, 作成の再現率が低いときには, 効果が無いが, 高い時には作成精度の向上に提案方式が貢献することを実験で示した.

また, プロトタイプの作成および, それを用いた単語の配置の補正の仕方にさらなる工夫が必要であることも明らかとなった. 今後は, それらについて検討を行い, その結果を情報検索や自然言語処理のタスクにおいても評価を行い, 単語のベクトル空間モデルの有効性を検証する予定である.

参考文献

[天野 00b] 天野, 近藤 : 日本語の語彙特性, 第 1 巻 単語親密度, 三省堂 (2000).

[別所 01] 別所克人 : 単語の概念ベクトルを用いたテキストセグメーテンション, 情報処理学会論文誌, Vol. 42, No. 11 (2001).

[Buckley 92] Buckley, C.: SMART version 11.0, ftp://ftp.cs.cornell.edu/pub/smart (1992).

[池原 97] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 (編) : 日本語語彙大系, 岩波書店 (1997).

[稲子 00] 稲子, 笠原, 松澤:複合語内単語共起による名詞の類似性判別方式, 情報処理学会論文誌, Vol. 41, No. 8 (2000).

[Hindle 90] Hindle, D.: Noun Classification from Predicate-Argument Structures, in *Proceedings of ACL*, pp. 268–275 (1990).

[笠原 97] 笠原, 松澤, 石川 : 国語辞書を利用した日常語の類似性判別, 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272–1284 (1997).

[笠原 03] 笠原, 稲子, 加藤 : テキストデータを用いた類義語の自動作成, 人工知能学会論文誌, Vol. 18, No. 4G 掲載予定 (2003).

[加藤 00] 加藤, 笠原, 北 : 概念検索に基づく技術内容からのエキスパートの推定, 信学技法, NLC2000 巻, pp. 55–62 (2000).

[金田一 88] 金田一, 池田 (編) : 学研 国語大辞典 第二版, 学習研究社 (1988).

[熊本 99] 熊本, 島田, 加藤 : 概念ベースの情報検索への適用—概念ベースを用いた検索特性の評価—, 情処研報, SIG-ICS-115, pp. 9 – 16 (1999).

[宮原 02] 宮原, 藤田, 安部, 林 : 散策型映像ポータルシステム AssociaGuide の提案, 電子情報通信学会総合大会, No. D-8-7, p. 104 (2002).

[永森 00] 永森, 笠原, 松澤 : 概念ベース構築における表記と概念のマッピング手法, 人工知能学会全国大会, 第 14 巻, pp. 163 – 164 (2000).

[Rosch 75] Rosch, E. and Mervis, C. B.: Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7, 573- 605 (1975).

[Schutze 92] Schütze, H.: Dimensions of Meaning, in *Proc. of Supercomputing 92*, pp. 787–796 (1992).

[Schutze 95] Schütze, H. and Pedersen, J.: Information retrieval based on word senses, in *Fourth Annual Sympo. on Document Analysis and Information Retrieval*, pp. 161–175 (1995).