6Q-1

Geometric Algebra を用いた英語文書分類手法の日本語文書への適用に関する問題についての基礎的検討

鈴木　直人†　　古橋　武†　　吉川　大弘†

名古屋大学†

1. 概要

電子文書の普及が進み，様々な場面で膨大な量の文書を管理する必要が生じている．このような文書管理においては，文書分類が不可欠となる．これまで，tf-idf や潜在意味解析(LSA)を用いた文書分類手法が報告されている[1]が，これらの多くは単語の出現順序を考慮していない．これに対し，英語文書において，Geometric Algebra (GA) [2]を用いることで，単語の出現順序を考慮して文書分類を行う手法が提案されている[3]．この手法では，LSA に基づき，単語の出現順序に応じた不可逆な回転ベクトルを定義することで，各文書が出現する単語の順序に応じて回転され，その最終状態の違いにより文書分類を行っている．本稿では，この手法を日本語の文書分類に適用する際に生じる問題点などについて検討を行う．

2. Geometric Algebra(GA)

本節では Geometric Algebra(GA)[2]について説明する．GA は複素数の自然な拡張であり，$(p+q)$ 個の直交基底 $\{\mathbf{e}_1,\mathbf{e}_2,...,\mathbf{e}_p,\mathbf{e}_{p+1},...,\mathbf{e}_{p+q}\}$ を持つ．この GA 空間は $\mathcal{G}(p,q)$ で表され，各基底は，$\mathbf{e}_i^2=1\,(i=1,2,...,p),\ \mathbf{e}_i^2=-1\,(i=p+1,...,p+q)$ を満たすように定義される．基底どうしの内積は，$\mathbf{e}_i\cdot\mathbf{e}_i=\mathbf{e}_i^2,\ \mathbf{e}_i\cdot\mathbf{e}_j=0\,(i\neq j)$ を満たすように定義される．

ここで，GA の特徴的な演算である GA 積について説明する．GA 積は，$\mathbf{e}_i\mathbf{e}_i=\mathbf{e}_i\cdot\mathbf{e}_i$，$\mathbf{e}_i\mathbf{e}_j=\mathbf{e}_{ij}\,(i\neq j)$ を満たすように定義され，$i\neq j$ の場合については外積に等しい($\mathbf{e}_i\mathbf{e}_j=\mathbf{e}_i\wedge\mathbf{e}_j\,(i\neq j)$)．また，GA 積は非可換な演算である($\mathbf{uv}=\mathbf{u}\cdot\mathbf{v}+\mathbf{u}\wedge\mathbf{v}$)．このような性質から，GA 積は内積と外積の和として表される($\mathbf{e}_i\mathbf{e}_j=\mathbf{e}_i\cdot\mathbf{e}_j+\mathbf{e}_i\wedge\mathbf{e}_j\,(i\neq j)$)．

次に，基底のグレードについて説明する．$\mathcal{G}(p,q)$ が持つ $(p+q)$ 個の直交基底 $\{\mathbf{e}_1,\mathbf{e}_2,...,\mathbf{e}_p,\mathbf{e}_{p+1},...,\mathbf{e}_{p+q}\}$ はすべてグレード 1 である．このグレード 1 の基底から任意の 2 つの基底 $\mathbf{e}_i,\mathbf{e}_j\,(i\neq j)$ を選び，その GA 積である $\mathbf{e}_{ij}$ をグレード 2 と呼ぶ．このように，基底の添え数字の個数，つまりグレード数は，グレード 1 の基底の GA 積の回数となる．また，定数の項のグレードは 0 とし，グレード n のベクトルは，n-ベクトルと呼ぶ．

具体例として $\mathcal{G}(3,0)$の場合で説明する．この GA 空間は，3 個の直交基底$\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$をもち，これらの基底の GA 積で生じ得るすべての基底の集合は正規基底と呼ばれ，$\{1,\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3,\mathbf{e}_{12},\mathbf{e}_{23},\mathbf{e}_{31},\mathbf{e}_{123}\}$ となる．このうち，$\{1\}$ はグレード 0 の基底，$\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$はグレード 1 の基底，$\{\mathbf{e}_{12},\mathbf{e}_{23},\mathbf{e}_{31}\}$はグレード 2 の基底，$\{\mathbf{e}_{123}\}$はグレード 3 の基底である．任意の GA 空間 $\mathcal{G}$(p,q)において，正規基底の個数は $2^{(p+q)}$ 個である．

次に，本稿で用いるベクトルの回転について説明する．ある 2 つの単位 1-ベクトル $\mathbf{u},\mathbf{v}$ について，この 2 つのベクトルがなす角を $\theta/2$ とする．$\mathbf{R}=\mathbf{uv},\mathbf{R}^\dagger=\mathbf{vu}$ とすると，任意のベクトル $\mathbf{a}$ に対して，$\mathbf{a}\mapsto\mathbf{RaR}^\dagger$ は，$\mathbf{a}$ を θ だけ回転させる演算である．以下では，この演算における $\mathbf{R}$ を回転ベクトルと呼ぶ．ここで，$\mathbf{B}=\dfrac{\mathbf{v}\wedge\mathbf{u}}{\|\mathbf{v}\wedge\mathbf{u}\|}$ とおくと，$\mathbf{R}=\mathbf{u}\cdot\mathbf{v}+\mathbf{u}\wedge\mathbf{v}$ は，$\mathbf{R}=\mathbf{u}\cdot\mathbf{v}-\|\mathbf{v}\wedge\mathbf{u}\|\mathbf{B}$ と書ける．$\mathbf{u},\mathbf{v}$ がともに 1-ベクトルであるため，$\mathbf{B}=\dfrac{\mathbf{v}\wedge\mathbf{u}}{\|\mathbf{v}\wedge\mathbf{u}\|}$ は単位 2-ベクトルとなり，$\mathbf{B}^2=-1$ である．そのため，$\cos\theta=\mathbf{u}\cdot\mathbf{v},\sin\theta=\|\mathbf{v}\wedge\mathbf{u}\|$ とおけば，$\mathbf{R}=\exp\left(-\mathbf{B}\dfrac{\theta}{2}\right),\mathbf{R}^\dagger=\exp\left(\mathbf{B}\dfrac{\theta}{2}\right)$ と書ける．

A basic study on application of English Document classification method using Geometric Algebra to Japanese Document Classification

†Naoto Suzuki, Takeshi Furuhashi, Tomohiro Yoshikawa, Nagoya University

3. GA を用いた文のベクトル化

本節では，GA を用いた文のベクトル化の手法について説明する．英語文書においては，GA を用いた文のベクトル化の手法が G. Pilato らによって提案されている[3]．この手法では，初めに，対象となる文の集合の全ての文書から，単語と単語の共起関係を表す単語-単語行列 $\mathbf{A}$ を作成する．この行列について，特異値分解(SVD)を用いて，低ランク(ランク k)の行列 $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$ で近似する．この特異値分解において，$\mathbf{U}_k$ 行列の i 行目の行ベクトル $\mathbf{l}_i$ に，$\mathbf{\Sigma}_k$ 行列の対角成分の i 番目の要素の平方根を掛けたベクトルは，i 行目に対応する単語 w_i の左側の文脈情報(文章において単語 w_i より前に出現する単語情報)を表す．一方，$\mathbf{V}_k$ 行列の i 列目の列ベクトル $\mathbf{r}_i$ に，$\mathbf{\Sigma}_k$ 行列の対角成分の i 番目の要素の平方根を掛けたベクトルは，単語 w_i の右側の文脈情報を表す．$\mathbf{R}'_{ij} = \mathbf{r}_i \mathbf{l}_j = \exp(-\mathbf{B}\theta)$ としたとき，回転ベクトル $\mathbf{R}_{ij} = \exp\left(-\mathbf{B}\frac{\theta}{2}\right)$ は，単語 w_i の後に，単語 w_j が出現する Bi-gram[4]に関連している．

各文を表すベクトルについて，初めに，ベクトル長 k，要素をすべて 1 とした 1-ベクトルを初期ベクトルとして作成する．次に，その文の単語の出現する順番に応じて，対応する回転ベクトル(例えば，単語 w_i - w_j の順で出現したとき，$\mathbf{R}_{ij}$)による回転の演算を行う．こうして，単語の出現順序を考慮した，各文を表すベクトルの最終状態が得られる．各文を表すベクトル $\mathbf{v}_i, \mathbf{v}_j$ の類似度 $\mathrm{sim}(\mathbf{v}_i, \mathbf{v}_j)$ は，コサイン類似度を用いて定義され，$\mathrm{sim}(\mathbf{v}_i, \mathbf{v}_j) = \cos^2(\mathbf{v}_i, \mathbf{v}_j)$ $(\cos(\mathbf{v}_i, \mathbf{v}_j) \geq 0),\quad 0\,(\mathrm{otherwise})$ で求められる．

4. 実験

3 節で示した手法を用いて，表 1 に示すような文どうしの類似度を求めた．単語-単語行列 $\mathbf{A}$ を低ランクの行列 $\mathbf{A}_k$ で近似する際のランク $k = 4$ とした．また，文の形態素解析には，Mecab(ver.0.994)を使用した．

表 2 から，文 A, B, C と文 D, E, F のそれぞれ 3 つの文どうしは類似度が 1 となっており，文集合のベクトル空間において，同じベクトルとみなされることがわかる．一方で，文 A, B, C と文 D, E, F では，類似度が 0.25 となっていることがわかる．単語の出現順序を考慮することで，従来の tf-idf や LSA などでは同一文として分類される A と D や B と E などが，異なるベクトルとして扱われることが確認できた．

表 1　実験に用いた文のリスト

A. 猫が鼠を見る
B. 猫が鼠を睨む
C. 猫が鼠を追う
D. 鼠が猫を見る
E. 鼠が猫を睨む
F. 鼠が猫を追う

表 2　文の類似度(k=4)

	A	B	C	D	E	F
A	1.00	1.00	1.00	0.25	0.25	0.25
B	1.00	1.00	1.00	0.25	0.25	0.25
C	1.00	1.00	1.00	0.25	0.25	0.25
D	0.25	0.25	0.25	1.00	1.00	1.00
E	0.25	0.25	0.25	1.00	1.00	1.00
F	0.25	0.25	0.25	1.00	1.00	1.00

5. おわりに

本稿では，Geometric Algebra を用いた英語文書での文のベクトル化の手法について説明した．また，日本語の文に対して適用し，単語の出現順序を考慮した分類が行えることを確認した．今後の課題としては，低ランクの近似行列 $\mathbf{A}_k$ を求める際のランク数 k に対する検討が挙げられる．また，文書における各文のベクトル化の用い方に対する検討も行う必要がある．

6. 参考文献

[1] F. Sebastiani, Machine learning in automated text categorization, ACM computing surveys, Vol.34(1), pp. 1-47, 2002

[2] D. Hestenes, New foundations for classical mechanics, Dordrecht, 1986

[3] G. Pilato, A. Augello, G. Vassallo, S. Gaglio, Geometric Algebra Rotors for Sub-symbolic Coding of Natural Language Sentences, KES 2007 / WIRN 2007, Part I, LNAI 4692, pp. 42-51, 2007

[4] C. E. Shannon, W. Weaver, R. E. Blahut, B. Hajek, The mathematical theory of Communication, University of Illinois press Urbana, 1949