動的計画法に基づく文脈の変化を考慮したLSAの検討*

◎佐古淳，滝口哲也，有木康雄 (神戸大)

1 はじめに

近年，音声認識手法の高精度化により，いくつかの場面で音声認識が用いられるようになってきた．しかし一方で，人間にとって不可解な認識誤りを起こしたり，単語列の認識から意味・内容の理解へどう進めるかといった問題も残されている．意味の通らない不可解な認識仮説を修正したり，認識結果の意味・内容を理解する上で，文章の semantic を考慮することは重要であると考えられる．現在，文章の semantic を考える上では Latent Semantic Analysis (LSA)[1] やこれを確率化した Probabilistic LSA (pLSA) などが用いられる．しかし，これらの手法は単語の出現順序を考慮しない Bag Of Words (BOW) に基づいており，また直接的には前後の文脈も考慮していない．より精細に semantic を扱うためには，文章の時系列も考慮に入れる必要があると考えられる．

本稿では，時系列を持った文書を対象とした LSA について考察を行う．Latent Semantic 空間の基底ベクトルを求める際，文書の主成分分析を行っていることに着目する．Kernel PCA [2] を用いて，高次元に射影された時系列文書の主成分分析を行う．このとき，正定値カーネルとして，動的計画法により計算される非線形時間伸縮が可能な Dynamic Time Alignment Kernel (DTA Kernel) [3] を用いることにより，時系列を考慮した Latent Semantic 空間を構築する．以下，基本的な LSA，提案手法，実験について述べる．

2 Latent Semantic Analysis

文書を表す特徴量として，単語頻度ベクトルが用いられるが，これは高次元でかつスパースなベクトルとなりやすい．LSA は，文書全体の集合から特異値分解を用いることで，文書をより低次元の Latent Semantic 空間へ射影する手法である．また，同時に単語の Latent Semantic 空間上での位置も得ることができる．LSA は以下のように特異値分解の形で定式化される．

$$\mathbf{W} = \mathbf{USV}^T. \tag{1}$$

ここで，$\mathbf{W}$ は，文書 d_j 中での単語 r_i の出現回数 w_{ij} を要素とする（語彙数 $M\times$ 文書数 N）の文書-単語共起行列，$\mathbf{U}$ は，$\mathbf{W}$ の基底となる（$M \times R$）の正規直交ベクトル，$\mathbf{S}$ は，対角に特異値を並べた，($R \times R$) の正方行列，$\mathbf{V^T}$ は，$\mathbf{W}$ の基底となる（$R \times N$）の正規直交ベクトル，$R = min(N, M)$ である．ただし，R は，次元圧縮のためより小さな値が設定される場合が多い．また，

$$\begin{aligned}\mathbf{W}^T\mathbf{W} &= \mathbf{VSU}^T\mathbf{USV}^T = \mathbf{VS}^2\mathbf{V}^T,\\ \mathbf{WW}^T &= \mathbf{USV}^T\mathbf{VSU}^T = \mathbf{US}^2\mathbf{U}^T,\end{aligned}$$

となることから，$\mathbf{V}$ は $\mathbf{W}^T\mathbf{W}$ の固有ベクトル，$\mathbf{U}$ は $\mathbf{WW}^T$ の固有ベクトルと考えられる．

3 提案手法

従来の LSA では，文書は $\mathbf{d}_j = (w_{1j} \cdots w_{Mj})^T$ というひとつのベクトルで表される．提案手法では，文書が時系列を持っている場合を考える．すなわち，$\mathbf{D}_j = (\mathbf{d}_j^{(1)} \cdots \mathbf{d}_j^{(T_j)})$ をひとつの文書と考える．ここで，$\mathbf{d}_j^{(t)} = (w_{1j}^{(t)} \cdots w_{Mj}^{(t)})^T$ は，単語の出現頻度ベクトル，T_j は文書に依存した値であり，文書によって長さが異なることを表す．文書の時系列を考慮する場合，比較するふたつの文書において，出現する単語が同じでも順序によって Latent Semantic 空間上で別の場所に配置されることが望ましい．しかし，従来の LSA では単語の順序が違い，意図が異なる文書であっても，単語の頻度が同じであれば同じ場所に射影されてしまう．提案手法では，このような問題を解決するため，順序を考慮した Latent Semantic 空間を構築する．

従来の LSA では，時系列を持った文書 $\mathbf{D}_j$ に対して Latent Semantic 空間を構築することはできない．ここで，Latent Semantic 空間の基底ベクトル $\mathbf{U}$ が，$\mathbf{WW}^T$ の固有値ベクトルとして与えられることに着目する．これは，$\mathbf{W}$ の主成分分析に等しい．このことから，Kernel PCA を用いて，時系列文書 $\mathbf{D}_j$ を高次元に射影し，高次元空間において主成分分析を行うことによって時系列文書に対する Latent Semantic 空間の構築を試みる．また，このとき正定値カーネルとして Dynamic Time Alignment (DTA) Kernel を用いることで，長さの違う時系列文書を扱うことが可能となる．すなわち，

$$\begin{aligned}K_{ij} &= \mathbf{\Phi}(\mathbf{D}_i) \cdot \mathbf{\Phi}(\mathbf{D}_j)\\ &= \max_{\phi_I, \phi_J} \frac{1}{L} \sum_{k=1}^{L} \mathbf{d}_i^{(\phi_I(k))} \cdot \mathbf{d}_j^{(\phi_J(k))}\end{aligned} \tag{2}$$

*Studies on Dynamic Programming Based LSA for Considering Context, by Atsushi SAKO, Tetsuya TAKIGUCHI and Yasuo ARIKI (Kobe University)

を用いる．ここで，$\mathbf{\Phi}(\mathbf{D}_i)$ は時系列文書 $\mathbf{D}_i$ の高次元空間での表現，ϕ_I, ϕ_J は時間伸縮関数，$L = max(T_j, T_i)$ である．K_{ij} は動的計画法によって計算することが出来る．$\bar{\mathbf{K}}$ の固有ベクトルを $\alpha^{(l)}$，固有値を λ_l とすると，Kernel PCA によって得られる第 l 主成分の基底ベクトル $\mathbf{u}_l$ は，

$$\mathbf{u}_l = \sum_{i=1}^{N} \hat{\alpha}_i^{(l)} \mathbf{\Phi}(\mathbf{D}_i) \quad (3)$$

として与えられる．ただし，$\hat{\alpha}^{(l)} = \alpha^{(l)}/\sqrt{N \cdot \lambda_l}$，

$$\bar{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N\mathbf{K} - \mathbf{K}\mathbf{1}_N + \mathbf{1}_N\mathbf{K}\mathbf{1}_N \quad (4)$$

であり，$\mathbf{K}$ は K_{ij} を要素とする行列，$\mathbf{1}_N$ は全ての要素が $1/N$ である $N \times N$ 行列である．文書 $\mathbf{D}_i$ の Latent Semantic 空間の軸 $\mathbf{u}_i$ への写像は，$\sqrt{\lambda_l} \cdot \alpha_i^{(l)}$ により得られる．

次に，未知の時系列文書 $\mathbf{X} = (\mathbf{x}^{(1)} \cdots \mathbf{x}^{(T_x)})$ を考える．$\mathbf{X}$ の Latent Semantic 空間の軸 $\mathbf{u}_l$ への写像は，

$$\mathbf{u}_l \cdot \mathbf{\Phi}(\mathbf{X}) = \sum_{i=1}^{N} \hat{\alpha}_i^{(l)} (\mathbf{\Phi}(\mathbf{D}_i) \cdot \mathbf{\Phi}(\mathbf{X})) \quad (5)$$

により得られる．これにより，未知文書とコーパス中の既知文書の関係を，時系列を考慮した Latent Semantic 空間で調べることが出来る．

4 実験

4.1 時系列文書に対する Latent Semantic 空間の構築

評価対象のコーパスとして，ラジオにおける野球実況中継の書き起こし文書を用いた．文書をまず，句点によって区切り，これを 1 文書の単位とした．さらに，1 つの文書に対して幅 ws の窓をかけ，窓中に含まれる単語の頻度ベクトルを抽出した．窓は $ws/2$ ずつずらし，時系列付きの単語頻度ベクトル集合を得た．全体の単語数は，約 8 万，文書数は約 9 千であった．この文書集合に対し，提案手法による Kernel PCA を行い，Latent Semantic 空間の基底ベクトルと，文書集合の写像を得た．

4.2 パープレキシティによる評価

提案手法を用いた Latent Semantic 空間の評価として，LSA に基づく言語モデルを構築し，パープレキシティを求めた．LSA に基づく言語モデルとして，

$$P(x_i|\mathbf{H}_{i-1}) \approx P(x_i|x_{i-1} \cdots x_{i-N+1}) P(\mathbf{v}_{i-1}|x_i)$$

を用いる．ただし，r_i は未知文書の i 番目の単語，$\mathbf{H}_{i-1} = (r_1, \cdots, r_{i-1})$，$\mathbf{v}_{i-1}$ は，$i-1$ 番目までの単語列の Latent Semantic 空間への写像である．ただし，$\mathbf{v}_{i-1}$ を求める際には，単語列に対して幅 ws の窓をかけることにより時系列付きの単語頻度ベクトルへ変換し，その後，Latent Semantic 空間への写像を求めた．また，$P(\mathbf{v}_{i-1}|x_i)$ は近似的に，$\mathbf{v}_{i-1}$ の近傍の文書に含まれる単語により unigram 確率のスケーリングした値を用いた．

学習データに対しオープンのテストセットから，テストセットパープレキシティを求めた．テストセットの単語数は約 2 万，文書数は約 2 千であった．結果を図 1 に示す．通常の LSA を用いた場合でも，n-gram よりパープレキシティは低下する．提案手法を用いた場合，$ws = 8$ のとき最もパープレキシティが低下しており，通常の LSA よりも若干の改善が見られた．1 文書は多い場合でも 30 単語程度から構成されていることから，ws が大きい場合の結果は LSA と同等となった．

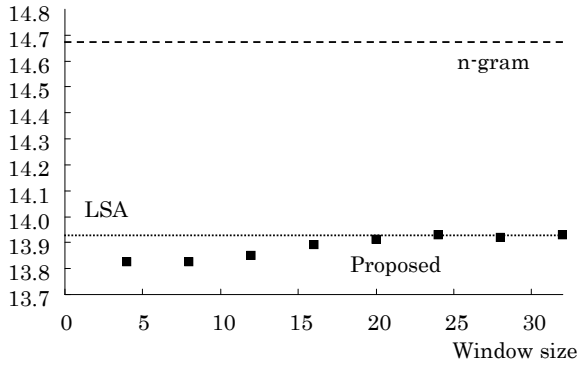

Fig. 1 Perplexity results.

5 おわりに

本稿では，時系列を持った文書を Dynamic Time Alignment (DTA) Kernel を用いた Kernel PCA によって Latent Semantic 空間に写像する手法について検討を行った．実験の結果，テストセット・パープレキシティにおいて若干の改善が見られた．

今後は，LSA と pLSA の関係のように，DTA Kernel を用いた Kernel PCA の確率的解釈について検討を行いたい．また，音声認識結果のクラス分類を行う上での提案手法の利用方法の検討行う予定である．

参考文献

[1] S. Deerwester *et al.*, "Indexing by latent semantic analysis," 1990.
[2] B. Schölkopf, *et al.*, Neural Computation, Vol.10, pp.1299–1319, 1998.
[3] 野間ら, PRMU, Vol.100, No.507, pp.63–68.