社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74 (43)
2008 / 12 / 10

単語出現順序を考慮したトピックモデルによる言語モデル適応

佐古　淳†　滝口 哲也††　有木 康雄††

† 神戸大学大学院自然科学研究科
〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
† 神戸大学大学院工学研究科
〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
E-mail: †sakoats@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし 人間にとって不可解な認識誤りの低減や，単語の認識だけでなく，意味・内容の理解を行うためには semantics を考慮することが重要であると考えられる．現在，LSA や PLSA のように semantics を考慮するモデルは Bag-of-words に基づく手法であり，文書中の単語出現順序を考慮していない．より高度な分析のためには，文書中の単語出現順序を考慮する必要があると考えられる．本研究では，Kernel PCA 及び Dynamic Time Alignment カーネルを用いることにより，単語順序を考慮した Latent Semantic 空間を構築する手法を提案する．予備実験では，右回り/左回りにプロットした時系列データが Latent Semantic 空間においてきれいに分離されることを確認した．また，言語コーパスを用いた評価実験では，パープレキシティの低下を確認することが出来た．

キーワード Latent Semantic Analysis, Kernel PCA, トピックモデル

Language Model Adaptation by Topic Model Based on Sequence of Words

Atsushi SAKO†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Guraduate School of Science and Technology, Kobe University
1–1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657–8501 Japan
† Guraduate School of Engineering, Kobe University
1–1 Rokkodaicho, Nada-ku, Kobe, Hyogo, 657–8501 Japan
E-mail: †sakoats@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract It is important to consider semantics for reductions of recognition errors unlike humans or understanding meanings and contents. To accommodate these problems, Latent Semantic Analysis (LSA) or Probabilistic LSA have been proposed. However these methods are based on Bag-of-words techniques. For more sophisticated analysis, it needs to consider a sequence of words in a document. In this paper, we propose the method based on Kernel PCA and Dynamic Time Alignment Kernel in order to consider a sequence of words. Preliminary experimental results shows the proposed method can separete clearly a sequence of right turn/left turn prots data. Moreover, experimental results of language corpus shows the reduction of perplexity.

Key words Latent Semantic Analysis, Kernel PCA, Topic Model

1. はじめに

近年，マルチメディア・コンテンツの増大に伴い，検索を容易に行うためのメタデータの需要が高まっている．メタデータの作成において，音声データからテキスト情報を抽出可能な音声認識技術が有用であると考えられる．現在，Web 上の Podcast を音声認識することでテキスト検索を可能としている PodCastle [1] や，MIT 講義音声を対象として MIT Lecture Browser [2]，音声認識と統計翻訳を組み合わせた多言語メディアブラウザ [3] などが提案されている．

しかし一方で，人間にとって不可解な認識誤りを起こすという問題や，単語列の認識からより高度なメタデータ作成のための意味・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 semantics 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) [6] やこれを確率化した Probabilistic LSA (pLSA) などが用いられる．また，pLSA のトピック分布を HMM の出力とすることで，文書間の文脈変化を考慮したモデル [4] や，より直接的にディリクレ分布を用いてトピック分布の変化を追跡するモデル [5] なども提案されている．

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，文書間での話題変化を考慮しているモデルにおいても，文書内での話題変化は考慮出来ていない．より精細に semantics を扱うためには，文章内の単語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，文書中の単語の時系列を保ったままでの LSA について考察を行う．しかしながら，通常の LSA では，全ての文書の特徴ベクトルが同じ次元数である必要がある．文書中の単語の時系列を考慮する場合，次元数の違いが問題となる．また，たとえ次元数が同じであっても，同じテンポで単語が出現するわけではないことから，単語出現の非線形な時間の伸縮にも対応する必要がある．そこで本研究では，LSA において，Latent Semantic 空間の基底ベクトルを求める際，文書の主成分分析 (PCA) を行っていることに着目する．通常の PCA を Kernel PCA [7] に置き換えることで，高次元に射影された主成分分析を行う．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) [8] を用いることにより，次元数の異なる文書への対応，及び非線形時間伸縮に対応する．

本稿では，以下，次節において基本的な LSA について述べる．その後，3. 節において Kernel PCA について述べ，4. 節において提案手法について述べる．5. 節において評価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，

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

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. Kernel PCA

ここで，d 次元観測ベクトルを $\mathbf{x}_j$ (j はフレーム番号) とすると，共分散行列 C は，

$$C = \frac{1}{N}\sum_{j=1}^{N}\bar{\mathbf{\Phi}}(\mathbf{x}_j)\bar{\mathbf{\Phi}}(\mathbf{x}_j)^T \tag{2}$$

$$\bar{\mathbf{\Phi}}(\mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_j) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{\Phi}(\mathbf{x}_j) \tag{3}$$

となる（N は全フレーム数）．C の固有値を λ，固有ベクトルを $\mathbf{v}$ と置くと，

$$\lambda\mathbf{v} = C\mathbf{v} \tag{4}$$

$$\lambda(\bar{\Phi}(\mathbf{x}_k)\cdot\mathbf{v}) = (\bar{\Phi}(\mathbf{x}_k)\cdot C\mathbf{v}),\quad k=1,\ldots,N \quad (5)$$

が得られる．また $\mathbf{v}$ は以下のようにサンプル点の線形結合で表現出来る．

$$\mathbf{v} = \sum_{i=1}^{N}\alpha_i\bar{\Phi}(\mathbf{x}_i) \quad (6)$$

式 (2) と (6) を (5) に代入すると左辺は，

$$\begin{aligned}\lambda(\bar{\Phi}(\mathbf{x}_k)\cdot\mathbf{v}) &= \lambda\sum_i \alpha_i\bar{\Phi}(\mathbf{x}_k)\cdot\bar{\Phi}(\mathbf{x}_i)\\ &= \lambda\sum_i\alpha_i\bar{K}_{ki} \quad (7)\end{aligned}$$

となる．ここで，

$$\bar{K}_{ki} = \bar{\Phi}(\mathbf{x}_k)\cdot\bar{\Phi}(\mathbf{x}_i) \quad (8)$$

とした．また右辺は，

$$\begin{aligned}&\bar{\Phi}(\mathbf{x}_k)\cdot C\mathbf{v}\\ &= \bar{\Phi}(\mathbf{x}_k)\cdot\frac{1}{N}\sum_j\bar{\Phi}(\mathbf{x}_j)\bar{\Phi}(\mathbf{x}_j)^T\sum_i\alpha_i\bar{\Phi}(\mathbf{x}_i)\\ &= \bar{\Phi}(\mathbf{x}_k)\cdot\frac{1}{N}\sum_i\alpha_i\left\{\sum_j\bar{\Phi}(\mathbf{x}_j)\bar{\Phi}(\mathbf{x}_j)^T\bar{\Phi}(\mathbf{x}_i)\right\}\\ &= \frac{1}{N}\sum_i\alpha_i\left[\bar{\Phi}(\mathbf{x}_k)\cdot\left\{\sum_j\bar{\Phi}(\mathbf{x}_j)\bar{\Phi}(\mathbf{x}_j)^T\bar{\Phi}(\mathbf{x}_i)\right\}\right]\\ &= \frac{1}{N}\sum_i\alpha_i\sum_j\left\{\bar{\Phi}(\mathbf{x}_k)\cdot\bar{\Phi}(\mathbf{x}_j)\right\}\left\{\bar{\Phi}(\mathbf{x}_j)\cdot\bar{\Phi}(\mathbf{x}_i)\right\}\\ &= \frac{1}{N}\sum_i\alpha_i\sum_j\bar{K}_{kj}\bar{K}_{ji} \quad (9)\end{aligned}$$

となる．従って，式 (7) と (9) より，

$$\begin{aligned}N\lambda\alpha &= \bar{\mathbf{K}}\alpha\\ \hat{\lambda}\alpha &= \bar{\mathbf{K}}\alpha \quad (10)\end{aligned}$$

となり，最終的に $\bar{\mathbf{K}}$ の固有値問題に帰着する事になる．ここで $N\lambda$ を $\hat{\lambda}$ とし，また $\bar{K}_{ki}$ を要素とする行列を $\bar{\mathbf{K}}$ とした．ただし，以下に示すように $\bar{K}_{ij}$ は K_{ij} から計算することが可能である．

$$\begin{aligned}\bar{K}_{ij} &= \bar{\Phi}(\mathbf{x}_i)\cdot\bar{\Phi}(\mathbf{x}_j)\\ &= \left(\Phi(\mathbf{x}_i)-\frac{1}{N}\sum_{m=1}^{N}\Phi(\mathbf{x}_m)\right)\\ &\quad\cdot\left(\Phi(\mathbf{x}_j)-\frac{1}{N}\sum_{n=1}^{N}\Phi(\mathbf{x}_n)\right)\\ &= \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)-\frac{1}{N}\sum_{m=1}\Phi(\mathbf{x}_m)\cdot\Phi(\mathbf{x}_j)\\ &\quad-\frac{1}{N}\sum_{n=1}\Phi(\mathbf{x}_n)\cdot\Phi(\mathbf{x}_i)\\ &\quad+\frac{1}{N^2}\sum_{m,n=1}\Phi(\mathbf{x}_m)\cdot\Phi(\mathbf{x}_n)\\ &= K_{ij}-\frac{1}{N}\sum_{m=1}1_{im}K_{mj}-\frac{1}{N}\sum_{n=1}K_{in}1_{nj}\\ &\quad+\frac{1}{N^2}\sum_{m,n=1}1_{im}K_{mn}1_{nj} \quad (11)\end{aligned}$$

$$K_{ij} = \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j) \quad (12)$$

$$1_{ij} = 1 \quad \text{for all } i,j \quad (13)$$

よって，$\bar{K}_{ij}$ の行列表現は次式で与えられる．

$$\bar{\mathbf{K}} = \mathbf{K}-\mathbf{1}_N\mathbf{K}-\mathbf{K}\mathbf{1}_N+\mathbf{1}_N\mathbf{K}\mathbf{1}_N \quad (14)$$

$\mathbf{1}_N$ は全ての要素が $1/N$ である $N\times N$ 行列である．

ここで，固有値を $\lambda_1\leqq\lambda_2\leqq\cdots\leqq\lambda_N$ とし，それに対応する固有ベクトルを $\alpha^{(1)},\cdots,\alpha^{(N)}$ とした際，

$$\mathbf{v}^{(l)}\cdot\mathbf{v}^{(l)} = 1,\quad \text{for all } l=p,\cdots,N \quad (15)$$

を満たすように，α を正規化する．(p 番目の固有値が，正の固有値の中で一番小さい値とする．) 式 (6) と (10) より，式 (15) は

$$\begin{aligned}1 &= \sum_{i,j}^{N}\alpha_i^{(l)}\alpha_j^{(l)}(\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j))\\ &= \sum_{i,j}^{N}\alpha_i^{(l)}\alpha_j^{(l)}K_{ij}\\ &= (\alpha^{(l)}\cdot\bar{\mathbf{K}}\alpha^{(l)})\\ &= \hat{\lambda}_l(\alpha^{(l)}\cdot\alpha^{(l)}) \quad (16)\end{aligned}$$

となる．よって，$\bar{\mathbf{K}}$ の固有ベクトル α に対して次式のように正規化を行う．

$$\hat{\alpha}^{(l)} = \frac{\alpha^{(l)}}{\sqrt{\hat{\lambda}_l}} \quad (17)$$

次に，高次元空間において主成分を抽出するために，テストデータ $\mathbf{y}$ の高次元における値 $\bar{\Phi}(\mathbf{y})$ を，固有ベクトル $\mathbf{v}^{(l)}$ 上に写像する．

$$\begin{aligned}(\mathbf{v}^{(l)}\cdot\bar{\Phi}(\mathbf{y})) &= \sum_{i=1}^{N}\hat{\alpha}_i^{(l)}(\bar{\Phi}(\mathbf{x}_i)\cdot\bar{\Phi}(\mathbf{y}))\\ &= \sum_{i=1}^{N}\hat{\alpha}_i^{(l)}\bar{K}^{test}(\mathbf{x}_i,\mathbf{y}) \quad (18)\end{aligned}$$

ここで，テストデータ $\mathbf{y}$ と学習データとの内積 K_{lj}^{test} を計算する．

$$K_{lj}^{test} = \Phi(\mathbf{y}_l)\cdot\Phi(\mathbf{x}_j) \quad (19)$$

式 (18) における $\bar{K}^{test}$ は，K^{test} から求めることが

出来る．

$$\bar{K}_{ij}^{test} = \left(\Phi(\mathbf{y}_i) - \frac{1}{N}\sum_{m=1}^{N}\Phi(\mathbf{x}_m)\right) \cdot \left(\Phi(\mathbf{x}_j) - \frac{1}{N}\sum_{n=1}^{N}\Phi(\mathbf{x}_n)\right) \quad (20)$$

$$\bar{\mathbf{K}}^{test} = \mathbf{K}^{test} - \mathbf{1}'_N\mathbf{K} - \mathbf{K}^{test}\mathbf{1}_N + \mathbf{1}'_N\mathbf{K}\mathbf{1}_N \quad (21)$$

テストデータ $\mathbf{y}$ のフレーム数が L の場合，$\mathbf{1}'_N$ は要素が全て $1/N$ の $L \times N$ 行列となる．

4. 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，$\mathbf{W}\mathbf{W}^T$ 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，

$$K_{ij} = \Phi(\mathbf{D}_i)\cdot\Phi(\mathbf{D}_j) = \max_{\phi_I,\phi_J}\frac{1}{L}\sum_{k=1}^{L}\mathbf{d}_i^{(\phi_I(k))}\cdot\mathbf{d}_j^{(\phi_J(k))} \quad (22)$$

を用いる．ここで，$\Phi(\mathbf{D}_i)$ 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，

$$\mathbf{u}_l = \sum_{i=1}^{N}\hat{\alpha}_i^{(l)}\Phi(\mathbf{D}_i) \quad (23)$$

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

$$\bar{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N\mathbf{K} - \mathbf{K}\mathbf{1}_N + \mathbf{1}_N\mathbf{K}\mathbf{1}_N \quad (24)$$

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，

$$\mathbf{u}_l\cdot\Phi(\mathbf{X}) = \sum_{i=1}^{N}\hat{\alpha}_i^{(l)}(\Phi(\mathbf{D}_i)\cdot\Phi(\mathbf{X})) \quad (25)$$

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．

5. 実　　験

本節では，提案手法による評価実験について述べる．まず始めに簡単な二次元データを用いた予備実験について述べ，提案手法の有効性を検証する．その後，実際の言語コーパスを利用し，パープレキシティにおいて評価する．

5.1 予備実験

本節では，提案手法を用いた予備実験について述べる．言語コーパスは特徴量の次元数も高く，また，単語の順序なども複雑なことから，提案手法の着眼点である順序とKernel PCA語の空間の関係について考察を行いにくい面がある．そのために，図1のようなシンプルなデータを用意し，提案手法により空間変換を行った．図中のデータは，それぞれ原点を中心に右回り/左回りのどちらかに数点ずつプロットされたものである．プロットの点数は固定ではなく乱数によってばらつきを与えた．また，座標についても同様に乱数を与えた．図の通り，右回りの点も左回りの点も，座標としては同じような場所に存在することがわかる．

提案手法を用いて空間変換を行った結果を図2に示す．図の通り，元の座標では同じような位置に存在していた右回り/左回りのシーケンスがきれいに

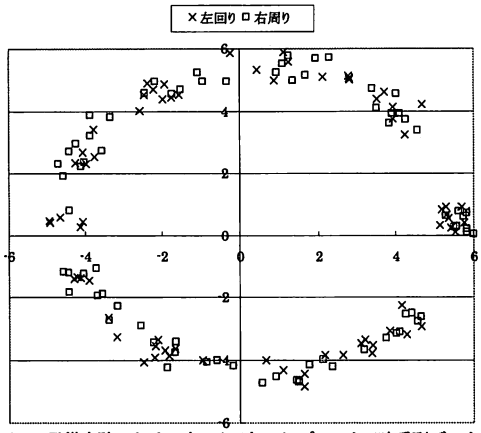

図 1　予備実験のための右回り/左回りプロット (時系列データ).

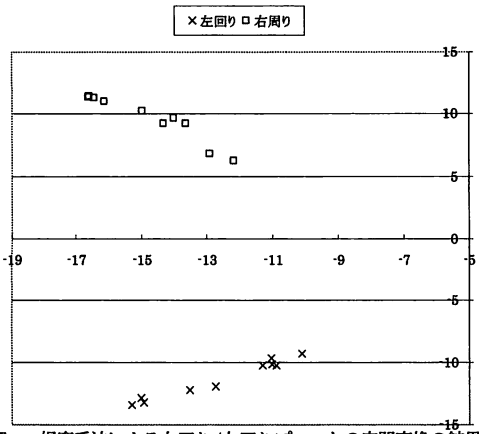

図 2　提案手法による右回り/左回りプロットの空間変換の結果.

分離していることがわかる．このことから，提案手法においては，元データの座標（頻度）が類似している場合でも，出現順序が異なれば遠い Latent Semantic 空間に射影されることが確認出来た．

5.2 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．

5.3 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，

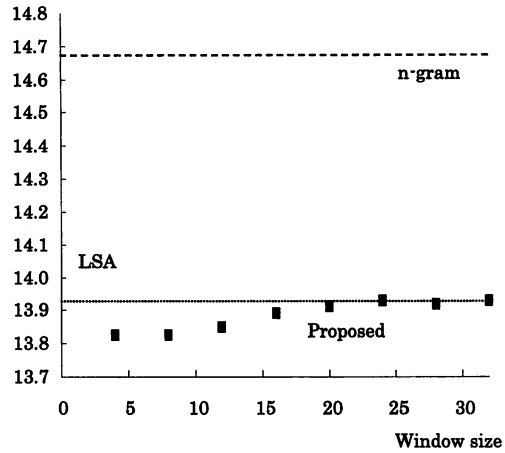

図 3　Perplexity results.

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

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 3 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．

6. ま と め

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．予備実験では，右回り/左回りにプロットしたシーケンスが，提案手法によってきれ

いに分離した空間に射影されることが確かめられた．また，言語コーパスを用いた実験の結果，テストセット・パープレキシティにおいて改善を確認することが出来た．文書中での単語の出現順序を詳細に考慮することで，モデルの性能を高めることが出来たと考えられる．

今後の課題として，単語の出現順序が違うからといって必ずしもLatent Semantic空間での距離を遠くすべきとは限らないことから，このような場合の対処を考える必要がある．また，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．

文　献

[1] M. Goto, J. Ogata, and K. Eto, "PodCastle: A Web 2.0 Approach to Speech Recognition Research," Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), pp.2397-2400, August 2007.

[2] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," INTERSPEECH 2007, pp. 2553–2556 (2007).

[3] 堀 貴明, 須藤 克仁, 大庭 隆伸, 渡部 晋治, 小川 厚徳, 渡辺 太郎, マクダーモット エリック, 塚田 元, 中村 篤, "「世界メディアブラウザ」 ー音声認識と統計翻訳に基づく多言語動画コンテンツ検索／閲覧システム," 日本音響学会講演論文集 1-1-17 (2008.9).

[4] A. Sako, T. Takiguchi, Y. Ariki, "Language Modeling Using PLSA-Based Topic HMM," IEICE TRANSACTIONS on Information and Systems, Vol.E91-D, No.3, pp.522-528, 2008.

[5] 岩田 具治, 渡部 晋治, 山田 武士, 上田修功, "トピックモデルに基づくユーザ興味の追跡, " 第 11 回情報論的学習理論ワークショップ (IBIS 2008), 2008.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, Landauer. T. K., and R. Harshman. "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, 41, 1990.

[7] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, Vol. 10, pp. 1299–1319 (1998).

[8] 野間 健一, 中井 満, 下平 博, 嶮峨山 茂樹, "非線形時間伸縮を用いた Support Vector Machine による時系列パタンの認識," 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, Vol.100, No.507(20001207) pp. 63-68, 2000.