DTA-Kernel PCA を用いた発話系列への意図ラベリングの検討*

◎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，音声認識が利用されつつある．しかし，現在の音声認識手法は，文法に基づく手法であり，発話の意図や意味といった semantics は考慮していない．そのため，人間にとって不自然な認識誤りを引き起こすといった問題がある．また，より高度なサービスのためには，単純な単語の認識よりも，意図・意味の理解が有用であると考えられる．本研究では，semantics を考慮した音声認識手法への一助として，発話系列への意図ラベリングを行った．

文書に対するラベリングにおいては，特徴量として，単語頻度ベクトルが用いられる．また，特異値分解に基づき単語頻度ベクトルをいくつかのトピックの軸に分解する Latent Semantic Analysis (LSA) や，これを確率化した Probabilistic LSA (pLSA) などが用いられる．これらの手法は，単語の出現順序を考慮しない Bag of Words (BOW) に基づいている．より精細に semantics 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．

本研究では，時系列を考慮した LSA 特徴として，正定値カーネルに Dynamic Time Alignment (DTA) Kernel を用いたカーネル主成分分析 (Kernel PCA) を用いる [1]．DTA-Kernel では，文書間の類似度を単語頻度ベクトルの内積ではなく，動的計画法 (DP) を用いて単語出現順序を考慮して計算する．これにより一発話についての特徴量が求まる．発話は単独ではなく，前後の発話との関連があると考えられる．そこで，本研究では，発話のコンテキストを考慮した手法として，意図ラベリングをチャンキングの問題と捉え，Conditional Random Fields (CRF)[2] によりラベル付けを行った．また，コンテキストを考慮しない場合の比較手法として，AdaBoost を用いて実験を行った．

2 DTA-Kernel PCA

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．DTA-Kernel PCA 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，$\mathbf{W}$ は，文書-単語共起行列としたとき，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，

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

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，

$$\mathbf{u}_l = \sum_{i=1}^{N} \hat{\alpha}_i^{(l)} \mathbf{\Phi}(\mathbf{D}_i) \quad (2)$$

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

$$\bar{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N \mathbf{K} - \mathbf{K}\mathbf{1}_N + \mathbf{1}_N \mathbf{K} \mathbf{1}_N \quad (3)$$

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．

---

*Studies on Intent Labeling for A Sequence of Utterances based on DTA-Kernel PCA, by Atsushi SAKO, Tetsuya TAKIGUCHI and Yasuo ARIKI (Kobe University)

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，

$$\mathbf{u}_l \cdot \mathbf{\Phi}(\mathbf{X}) \quad = \quad \sum_{i=1}^{N} \hat{\alpha}_i^{(l)} (\mathbf{\Phi}(\mathbf{D}_i) \cdot \mathbf{\Phi}(\mathbf{X})) \qquad (4)$$

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．

3 実験

本節では，DTA-Kernel PCA を用いた意図ラベリング実験について述べる．特徴量としては，DTA-Kernel PCA に加え，単語頻度ベクトル，LSA を用いた．特徴量はそれぞれ音声認識結果から求めた．また，識別器には，コンテキストを考慮した識別器として CRF を用いた．コンテキストを考慮しない識別器として AdaBoost を用いた．以下，コーパスの仕様，音声認識条件と結果，識別実験について述べる．

3.1 コーパスの仕様

実験のためのコーパスとして，ラジオの野球実況中継音声を用いた．これは，試合を伝えるという明確な目的が存在するため，意図ラベルの付与が比較的容易で明確なためである．データは 4 試合分で，それぞれ約 1.5 時間，2000 発話であった．まず，人手による書き起こしテキストを作成した．書き起こす際，主観により句点を挿入した．以後，実験では発話の単位として，この句点で区切られた一文を用いた．

教師データとして，発話毎に意図のラベルを人手で付与した．ラベルは大きくイベントと実況に分けた．イベントには，投球・ストライク・ボール・ファール・フォアボール・三振・牽制球・盗塁・ヒット・ツーベース・ホームラン・得点・アウト・ダブルプレー・デッドボールがある．また，実況には，実況一般・解説者との会話・守備の実況がある．

3.2 音声認識条件・結果

ベースラインの音響モデルは，日本語話し言葉コーパス (CSJ) モニター版のうち，男性話者 200 名の講演音声を用いて作成した．ベースラインの音響モデルを作成した後，MLLR+MAP により音響モデル適応を行った．音響モデル適応は，同一話者の別の日時の実況中継音声を用いて行った．適応データの分量は，約 1 時間半であった．

言語モデルは，野球実況中継音声の書き起こしテキストから trigram モデルを作成した．異なり単語数は約 3,000，コーパスサイズは約 8 万形態素であった．

音声認識の結果，単語正解精度は 4 試合全体で 63.8%であった．

3.3 識別実験

精度の指標としてイベント検出の F 値，及びイベントの正解率を用いる．特徴量は，単語頻度・DTA-Kernel PCA・頻度と DTA-Kernel PCA の組み合わせの 3 種類，識別器は，CRF，AdaBoost を用いた．F 値が最大になる場合の結果を表 1 に示す．実験の結果，特徴量に単語頻度+DTA Kernel PCA・識別器に CRF を用いた場合に，イベント検出・イベント正解率共に最もよい性能を示した．イベント検出は，どの特徴量・識別器においても，それぞれ 0.95 程度の高い性能を示した．一方で，イベント正解率は，単語頻度を組み合わせる方が高い性能が得られている．これは，コーパスの性質上，具体的なキーワードによってイベントの内容が決定される場合が多いためと考えられる．DTA-Kernel PCA では，全く具体的な単語を用いないにも関わらず，キーワードを用いた場合と近い性能を実現している．より多様で曖昧な表現を受理する必要がある場合に有効な手法となる可能性が示された．

Table 1　実験結果

イベント検出F値			
	単語頻度	KPCA	頻度+KPCA
CRF	0.96	0.94	0.97
AdaBoost	0.96	0.95	0.96
イベント正解率			
	単語頻度	KPCA	頻度+KPCA
CRF	0.82	0.78	0.84
AdaBoost	0.82	0.76	0.80

4 おわりに

本稿では，Dynamic Time Alignment (DTA) Kernel を用いた Kernel PCA によって単語の出現順序を考慮して文書の Latent Semantic を考慮した特徴量を抽出し，発話系列に対し意図のラベリングを行う手法について検討を行った．実験の結果，出現順序を考慮しない単語頻度を用いる場合に比べてイベント検出 F 値，イベント正解率共に向上した．また，識別器では，発話のコンテキストを考慮する CRF が最も良い精度となった．DTA-Kernel PCA を用いた手法は，多様な表現を受理できる可能性があると考えられる．

参考文献

[1] 佐古ら, 音講論, pp.221-222，2008-03.

[2] Lafferty, J., *et al.*, Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, pp.282–289, 2001.