周波数スペクトル分析を用いた人の動作に関わる副詞の意味理解

理学専攻・情報科学コース
2040649
谷口 巴

1 はじめに

本研究では，人の動作の特徴を通じて副詞の意味を理解することを試みる．具体的には，人の動作をガウス過程潜在変数モデル（GPLVM）[1] で圧縮して得られた非線形な潜在空間での軌跡を，スペクトル混合カーネル [2] を用いたガウス過程で表現する．さらに得られた各次元の軌跡を構成する複数の周波数成分と副詞との対応関係を捉える，周波数空間でのマルチモーダルなトピックモデルを提案する．

2 動作と副詞の結合トピックモデル

動作から抽出された周波数成分は, その動作に付与された副詞と関係があると考えられる．そこで, 潜在ディリクレ配分法 (LDA) [3] を拡張し, 各動作 d に K 次元の潜在的な「トピック分布」θ_d があると仮定する．このとき, 動作に付与された副詞 $\{w_{dn}\}$ $(n = 1 \dots N_d)$ および動作の周波数成分 $\{x_{dm}\}$ $(m = 1 \dots M_d)$ は, 次のモデルで生成されたと考える．

1. Draw $G_0 \sim \mathrm{DP}(\gamma, H)$.
2. Draw $\theta_d \sim \mathrm{DP}(\alpha, G_0)$.
3. For $n = 1 \dots N_d$,
   - Draw $z_{dn} \sim \theta_d$; Draw $w_{dn} \sim \mathrm{Mult}(\phi_{z_{dn}})$.
4. For $m = 1 \dots M_d$,
   - Draw $y_{dm} \sim \theta_d$;
     Draw $x_{dm} \sim N(\mu_{y_{dm}}, \Sigma_{y_{dm}})$.

このグラフィカルモデルを図 1に示した．

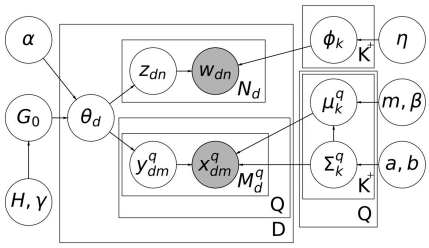

図 1: HDP-SMLDA のグラフィカルモデル.

このマルチモーダルなトピックモデルを, 本論文では階層的ディリクレ過程スペクトル混合 LDA (HDP Spectral Mixture LDA; HDP-SMLDA) と呼ぶ. HDP-SMLDA では、周波数成分と副詞は動作毎に同じトピック分布 θ_d を共有している．ここで $\phi_k, \mathcal{N}(\mu_k, \Sigma_k)$ はそれぞれ, k 番目のトピックに対応する副詞の多項分布および周波数のガウス分布であり, 互いの情報を用いて, 各動画について 1 つ 1 つの副詞と周波数成分にトピックを割り当てていく. また，中華料理店過程を階層的に導入することでトピック数をデータから自動的に推定する．

副詞と周波数についてのサンプリング　ギブスサンプリングにより, 副詞と周波数のトピック分布を学習していく．テーブルの割り当て集合を T, テーブルの卓番を l とすると，中華料理店過程に従い，副詞 w_{dn} のトピック z_{dn} は, 次式で座るテーブル t_{dn} をサンプリングすることで決定される．ここで，l_{used}, l_{new} はそれぞれ既存テーブルと新規テーブルを表し，L_k，L はそれぞれトピック k が振られたテーブル数，総テーブル数，V は語彙数を表す．

$$p(t_{dn} = l | \boldsymbol{W}, \boldsymbol{T_{\backslash dn}}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta)$$
$$\propto \begin{cases} (N_{dl\backslash dn} + \sum_{q=1}^{Q} M_{dl}^{q}) \frac{N_{kw}+\eta}{N_{k\backslash dn}+\eta \boldsymbol{V}} & (l = l_{used}) \\ \sum_{k=1}^{K} \frac{\alpha L_k}{L+\gamma} \frac{N_{kw}+\eta}{N_{k\backslash dn}+\eta \boldsymbol{V}} + \frac{\alpha\gamma}{L+\gamma} \frac{1}{V} & (l = l_{new}) \end{cases}$$

既存テーブルのには各々トピックが振り分けられており，新規テーブルがサンプリングされた際は，次式を用いてテーブルに置く料理 (トピック) を決定する．ここで，k_{used}, k_{new} はそれぞれ既存トピックと新規トピックを表す．

$$\theta_{dk}\phi_{kv}$$
$$\propto \begin{cases} (N_{dl\backslash dn} + \sum_{q=1}^{Q} M_{dl}^{q} + \frac{\alpha L_k}{L+\gamma}) \frac{N_{kv}+\eta}{N_k+\eta V} & (k = k_{used}) \\ \frac{\alpha\gamma}{L+\gamma} \frac{1}{V} & (k = k_{new}) \end{cases}$$

ハイパーパラメータである η は不動点反復法により，以下の式を用いて更新する．ここで登場する Ψ はディガンマ関数 $\Psi(x) = d/dx \log \Gamma(x)$ である．

$$\eta^{new} = \eta \frac{\sum_{k=1}^{K} \sum_{v=1}^{V} \Psi(N_{kv} + \eta) - KV\Psi(\eta)}{V \sum_{k=1}^{K} \Psi(N_k + \eta V) - KV\Psi(\eta V)} \tag{1}$$

周波数成分 x_{dm} のトピック y_{dm} に関しては, 副詞の単語分布をガウス分布の確率密度関数 f に置き換え，以下の式を用いてサンプリングする．

$$p(t_{dn} = l | \boldsymbol{W}, \boldsymbol{T_{\backslash dn}}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta)$$
$$\propto \begin{cases} (N_{dl\backslash dn} + \sum_{q=1}^{Q} M_{dl}^{q}) f(x|\mu_k, \Sigma_k) & (l = l_{used}) \\ \sum_{k=1}^{K} \frac{\alpha L_k}{L+\gamma} f(x|\mu_k, \Sigma_k) + \frac{\alpha\gamma}{L+\gamma} f(x|\mu_{k_{new}}, \Sigma_{k_{new}}) & (l = l_{new}) \end{cases}$$

$$\theta_{dk} f_{kx}$$
$$\propto \begin{cases} (N_{dl} + \sum_{q=1}^{Q} M_{dl\backslash dm}^{q} + \frac{\alpha L_k}{L+\gamma}) f(x|\mu_k, \Sigma_k) & (k = k_{used}) \\ \frac{\alpha\gamma}{L+\gamma} f(x|\mu_{k_{new}}, \Sigma_{k_{new}}) & (k = k_{new}) \end{cases}$$

パラメータである μ は以下の事後分布からサンプリングする．ここで $\lambda = 1/\sigma^2$ である．

$$p(\mu|\mathbf{Y}) = \mathrm{N}(\mu|m, (\beta\lambda)^{-1}) \tag{2}$$

ただし β_0, m_0 を事前分布のパラメータとして

$$\beta = M + \beta_0,\ m = \frac{1}{\beta} \left(\sum_{m=1}^{M} x_m + \beta_0 m_0 \right). \tag{3}$$

集中度 α の推定　集中度 α を推定することでよりデータにフィットしたトピック数を推定することがでする．α は以下の事後分布からサンプリングする．

$$\alpha \sim Ga(\alpha|c_1 + K^{+} - s, c_2 - log\pi) \tag{4}$$

ただし π, s を事前分布のパラメータとして

$$\pi \sim Beta(\pi|\alpha + 1, N + M),\ s \sim Bernoulli\Big(s|\frac{\frac{N+M}{\alpha}}{1 + \frac{N+M}{\alpha}}\Big). \tag{5}$$

3　実験

3.1　使用データ

YouTube に掲載されている, 100 種類の異なる歩行動作を集めた動画[1] を用いて実験を行った．クラウドソーシングシステム Lancers[2] を用いて，20 名のアノテーターに動画の各動作について思いつく限り自由に副詞をアノテーションしてもらうよう依頼した．全動画で 3 個以上出現した副詞に限定し，満たない副詞はノイズとして除去した．1 つの動画につき平均 12.93 個の副詞がアノテーションされたデータが得られた．

データの前処理　上記の動画データから，以下のように 4 段階で 3 次元の骨格座標の推定を行った．

1. Openpose [4] を用いて動画データから 2 次元の骨格座標を推定
2. FCRN-depth prediction [5] を用いて動画データの深度を推定
3. 1,2 の推定結果と 3d-pose baseline [6] を用いて動画データから 3 次元の骨格座標を推定
4. 歩いている人の体の向きを合わせるため, 回転行列を用いて正規化

周波数成分の抽出　前処理したデータから各関節間ごとに方向ベクトルを算出し, 入力データとした．以下の 2 つの手法を用いて, 前処理した動画データから周波数成分を抽出した．

1. GPLVM [1] を用いて 48 次元の姿勢データを 3 次元の潜在変数に圧縮する
2. SM kernel [2] を用いて 3 次元の潜在変数から, 各次元について周波数成分を抽出する

SM kernel では構成されるカーネル関数に関して重み $\mathbf{w}$, 平均 $\boldsymbol{\mu}$, 分散 $\boldsymbol{\nu}$ が推定されるが, イテレーションごとに重み $\mathbf{w}$ を用いて周波数成分平均 $\boldsymbol{\mu}$ をサンプリングすることとした．本実験ではこの抽出された周波数成分と動画に付与された副詞の集合を入力データとした．十分収束するよう, MCMC の繰り返し数は 1000 と大きい値に設定した．混合ガウス分布の部分に関して, 分散はデータの幅に合わせて K 個のガウス分布が均等に配置されるよう, σ=(周波数の範囲)/4K と固定して平均のみ学習する．

3.2　実験結果

推定されたトピック数は 23 であった．8 トピックについて学習されたトピック–単語分布から各副詞について NPMI [7] を計算した上位 3 語を表 1に示す．また学習された平均 $\boldsymbol{\mu}$ を用いてガウス分布を描画したものを図 2に示す．MCMC の各繰り返しごとに計算したパープレキシティを算出し収束していることを確認した．

[1] https://www.youtube.com/watch?v=HEoUhlesN9E
[2] https://www.lancers.jp/

表 1: HDP-SMLDA で得られたトピック別副詞上位 3 語.

Topic 1	Topic 2	Topic 3	Topic 4
急いで	堂々と	楽しそうに	ゆっくり
足早に	颯爽と	軽快に	ゆっくりと
素早く	力強く	リズミカルに	じっくりと
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
ふらふら	慎重に	痛そうに	踊るように
よろよろ	恐る恐る	ぎこちなく	面白く
よたよた	そろそろ	弱弱しく	コミカルに

3.3　モデルの評価

評価の指標としてパープレキシティを採用する．訓練前の単語分布を用いた周波数情報を考慮しないユニグラムモデルは 155 であった．またデータを訓練用と評価用に 9.5 対 0.5 に分割した. 同データを用いて訓練時の LDA では 99 まで低下した．副詞に加えて周波数情報を補助的に扱う HDP-SMLDA では訓練時に 52，評価時に 89 と LDA の訓練時のものより低い値であった．このことから，提案モデルが正しく副詞を予測していることが確認できた．

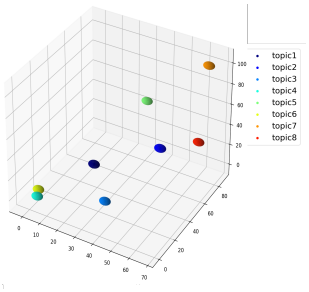

図 2: 学習した μ を用いて描画したガウス分布.

4　おわりに

歩行動作をする人間の骨格座標を GPLVM を用いて 3 次元の潜在変数に圧縮し, SM kernel を用いて周波数成分を抽出した. 次に HDP-SMLDA を用いて, 動画にアノテーションされた副詞データと周波数データを, お互いの情報を使いながら適切なトピック数にクラスタリングした. 最後にパープレキシティを用いて評価し, モデルの有効性を示した.

参考文献

[1] Michalis K. Titsias and Neil D. Lawrence. Bayesian Gaussian Process Latent Variable Model. In *AISTATS 2010*, pp. 844–851.

[2] Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian Process Kernels for Pattern Discovery and Extrapolation. In *ICML 2013*, pp. 1067–1075.

[3] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 994–1022, 2003.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 43, No. 1, pp. 172–186, 2021.

[5] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *3DV*, pp. 239–248, 2016.

[6] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A Simple yet Effective Baseline for 3D Human Pose Estimation. In *ICCV 2017*, pp. 2640–2649, 2017.

[7] Gerlof Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, pp. 31–40, 2009.