修士論文 要旨

単語埋め込み空間を用いた階層的トピックモデルの構成

数理情報学専攻 48186228　吉田 崇裕
指導教員　大西立顕 准教授

1　はじめに

本研究では，Gaussian LDA(GLDA)[3] の有する欠点である，トピック間の関係性を考慮できないこと及び多義語の文脈に応じた解釈が困難であることを改善すべく，階層的トピックモデルである hierarchical LDA(hLDA)[2] と GLDA を組み合わせたモデルである GhLDA を提案する．そして，提案手法はトピックの階層構造を抽出することに成功しただけでなく，GLDA ではできない「多義語の文脈に応じた意味解釈」が可能であることを示した．さらに，PMI(Pointwise mutual information) やテスト文書に対する対数尤度という定量的指標においても，既存手法に比べ高い性能を発揮することを示した．なお，修士論文には GhLDA の他に試みた手法 [4] についても記載しているところ，かかる手法によってはトピック間の階層構造を捉えることはできなかった．

2　既存研究

GLDA は，Latent Dirichlet Allocation(LDA)[1] の拡張モデルであり，学習済みの単語の分散表現を用い各単語をベクトルとして扱うのが大きな特徴である．また，hLDA も LDA の拡張モデルであり，トピックを木構造に配置し，文書全体を取りまとめるパスと各単語トークンごとに定まるレベルという 2 つの潜在変数から各単語トークンのトピックが定まる．また，木構造は nCRP(Nested Chinese Restaurant Process) によって定まり，可変である．

そして，GLDA におけるトピックという潜在変数の推論にはギブスサンプリングが用いられるが，そのサンプリング式は具体的に計算すると事実上「単語埋め込み空間上で混合ガウスモデルの推論をする」ものと同一であることが分かる．これは，文書内の他単語のトピックの影響が著しく弱いことを意味し，各単語のトピックが単純に単語埋め込み空間上どこに位置するかにのみ依存して決定されるから，GLDA は多義語の文脈に応じた意味解釈ができない．一方で文書ごとに取りうるトピックに制限を設ければ，文脈に応じた解釈が可能になると考えられる．さらに，GLDA はトピック間の関係性を考慮に入れておらず，トピックの階層構造を導入することで，よりコンパクトなトピック表現が可能になると期待される．

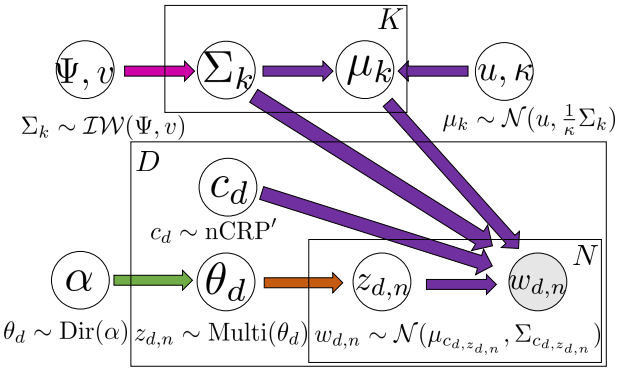

図 1. GhLDA-fixed のグラフィカルモデル

3　提案手法

本研究では，GLDA と hLDA を組み合わせた GhLDA のうち，トピックの木構造が固定であるものを GhLDA-fixed，可変であるものを GhLDA-BNP と呼ぶ．特に，GhLDA-fixed のモデリングは以下の通りである．まず，D を文書数，N_d を文書 d 内の単語トークン数，K をトピック数，θ_d を文書 d のトピック分布，c_d を文書 d に割りあてられたパス，$z_{d,n}, w_{d,n}$ を文書 d の n 番目の単語トークンに割りあてられたレベルと具体的な単語ベクトルとする．このとき，$\alpha, u, \kappa, \Psi, v$ をパラメータとして，

$$
\begin{aligned}
\mu_k, \Sigma_k &\sim \mathcal{NIW}(u, \kappa, \Psi, v)\ (k = 1, 2, \cdots), \\
c_d &\sim \mathrm{nCRP}', \\
\theta_d &\sim \mathrm{Dir}(\alpha)\ (d = 1, 2, \cdots, D), \\
z_{d,n} &\sim \mathrm{Multi}(\theta_d)\ (n = 1, 2, \cdots, N_d), \\
w_{d,n} &\sim \mathcal{N}(\mu_{c_{d,z_{d,n}}}, \Sigma_{c_{d,z_{d,n}}})\ (n = 1, 2, \cdots, N_d)
\end{aligned}
$$

として文書を生成する確率モデルが GhLDA-fixed である．GhLDA-fixed のグラフィカルモデルを図 1 に示す．ただし，nCRP′ とは，nCRP を簡易化した確率モデルであり，以下の性質を満たす．

$$
p(c_d | c_{-d}) = \frac{\#\{d' | c_{d'} = c_d, d' \neq d\}}{D - 1}.
$$

GhLDA-BNP も新たなトピックが出現しうる点を除けば fixed の場合と同様のモデリングである．また，GhLDA においては $c_d, z_{d,n}$ の事後確率を求めることが

		パスの学習結果		
		「銀行」	「土手」	その他
ラベル	「銀行」	**425**	3(→0)	39
	「土手」	0	**118**	4

表 1. GhLDA-fixed における “bank”の多義解釈性（「3(→0)」はラベリングミスのため実質的に 0 であることを表す.）

目標となるが，解析的に求めるのは困難であるためギブスサンプリングを用いて近似する．計算をすると，パス c_d のサンプリング式は

$$p(c_d|c_{-d},z,w)$$
$$=\prod_{l=1}^{L}\frac{1}{\pi^{|\mathcal{D}_{\mathrm{new},l}|M/2}}\frac{\Gamma_M\left(\frac{v_{\mathrm{all}}}{2}\right)}{\Gamma_M\left(\frac{v_{\mathrm{part}}}{2}\right)}\frac{|\Psi_{\mathrm{part}}|^{v_{\mathrm{part}}/2}}{|\Psi_{\mathrm{all}}|^{v_{\mathrm{all}}/2}}\left(\frac{\kappa_{\mathrm{part}}}{\kappa_{\mathrm{all}}}\right)^{\frac{M}{2}}$$
$$\cdot\frac{\#\{d'|c_{d'}=c_d\}}{D-1}$$

となり，レベル $z_{d,n}$ のサンプリング式は

$$p(z_{d,n}=l|c,z_{-(d,n)},w)$$
$$\propto\frac{\alpha_l+(N_{d(-n)})_l}{\sum_{l'}(\alpha_{l'}+(N_{d(-n)})_{l'})}$$
$$\cdot\mathcal{T}_{v_k-M+1}\left(w_{d,n}|u_k,\frac{\kappa_k+1}{\kappa_k(v_k-M+1)}\Psi_k\right)$$

となる．

4 実験結果

表 1 は，Wikipedia コーパスに存在する，“bank(s)”という多義語を含み “Rivers”，“Banks/Financial”のラベルが付された 589 文書について，GhLDA-fixed の多義語の意味解釈が可能であるかを実験した結果である．

表 1 によると，金融のラベルが付された文書に出現する “bank(s)”（「銀行」という意味で使われていると考えられる.）と河川のラベルが付された文書に出現する “bank(s)”（「土手」という意味で使われていると考えられる.）について，GhLDA は互いに混同することなく分類できていることが分かる．すなわち，GLDA ではできない多義語の文脈に応じた解釈が GhLDA においては可能であることが分かる．

また，各手法により学習したトピック分類結果を元に，トピックごとの上位 10 単語について PMI を計測し，その平均を取ったものを表 2 に示す．PMI は単語同士の意味的一貫性を表す指標であり，高い数値ほど同じトピック内の単語の意味的一貫性が高いことを示すから，既存手法に比べて GhLDA は同じトピックに意味一貫性の高い単語をまとめていることが分かる．

モデル	GloVe	word2vec	fasttext
LDA	-3.32		
hLDA	-1.06		
GLDA	-1.18	-2.20	-2.56
CGTM	-1.07	-1.63	-1.87
GhLDA-fixed	**-0.40**	-0.79	-0.96
GhLDA-BNP	-0.75	**-0.35**	**-0.77**

表 2. 各手法におけるトピック分類の意味一貫性（文書データ:Wikipedia）

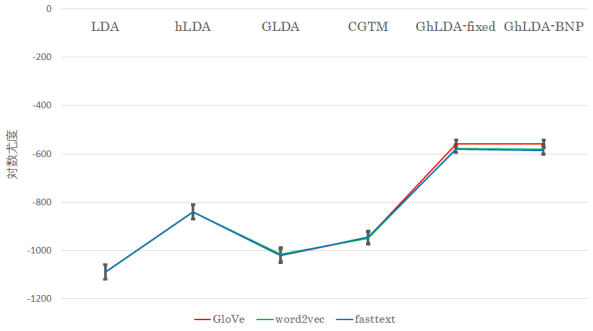

図 2. 各手法におけるテスト文書に対する対数尤度（文書データ:Wikipedia）

さらに，各手法により訓練データに対して学習した訓練データに対するトピック割当てを元に，テストデータに対する平均対数尤度を計算した結果を図 2 に示す．これによれば，テスト文書に対する対数尤度は，GhLDA で学習した場合で最も高くなることが分かる．

このように，GhLDA は多義語の文脈に応じた解釈ができることが定性的考察で明らかになったとともに，定量的評価においても GhLDA が既存手法より優れた性能を発揮することが分かった．

参考文献

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993—1022, Feb. 2003.

[2] D. M. Blei, and T. L. Griffiths, “The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies,” *Journal of the Association for Computing Machinery*, vol. 57, No. 2, Article 7, Jan. 2010.

[3] R. Das, M. Zaheer, and C. Dyer, “Gaussian LDA for Topic Models with Word Embeddings,” *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 795–804, Beijing, China, Jul. 2015.

[4] 吉田崇裕, 久野遼平, 大西立顕, “トピック間の階層構造を考慮した Gaussian LDA の構成,” 情報処理学会研究報告自然言語処理, 2019–NL–241 巻, 6 号, pp. 1–8, 2019 年.