情報処理学会論文誌　データベース　Vol. 2　No. 2　86–95 (June 2009)

学術文献の潜在トピックに着目したタンパク質相互関係に関する知識の抽出

麻生竜矢[†1]　江口浩二[†1]

近年，医学生物学分野をはじめとする様々な領域において，電子化された大量の文献に蓄積された知見を組織化し，潜在的な仮説を生成する技術への高度化への要求が高まっている．この目的の下で，生物学的知識，特にタンパク質間相互関係に関する知識の抽出のため，確率的トピックモデルを適用する．潜在的ディリクレ配分法（LDA）による確率的トピックモデルは，上述のようなタスクに関する有効性という観点からはこれまで検討されてこなかった．本論文では，LDA の推定手法として Collapsed 変分ベイズ法を適用し，テストセットの対数尤度，分類精度ならびにランキング精度の観点から評価し，Collapsed Gibbs Sampling 法による LDA と確率的潜在意味インデクシング法（pLSI）との比較を行う．現実的なタスクを想定した分類精度とランキング精度による評価では，Collapsed 変分ベイズ法に基づく LDA によって良好な結果が得られることを示す．

Extracting Knowledge on Protein-Protein Relationships Using Latent Topics of Literature

Tatsuya Aso[†1] and Koji Eguchi[†1]

Recently, technologies for organizing knowledge accumulated in a growing number of digitized documents and for generating potential hypotheses have been highly requested, such as in biomedical fields. For these objectives, we investigate applying statistical topic models to predict relationships between biological entities, especially protein mentions. A statistical topic model, Latent Dirichlet Allocation (LDA) has not been investigated for such a task. In this paper, we apply the state-of-the-art Collapsed Variational Bayesian inference to estimating the LDA model, and compared it with the LDA model estimated via Collapsed Gibbs Sampling and probabilistic Latent Semantic Indexing (pLSI) from the viewpoints of test-set log-likelihood, classification accuracy and ranking effectiveness. We demonstrate through experiments that the Collapsed Variational LDA gives better results than the others, especially in terms of classification accuracy and ranking effectiveness.

1. はじめに

近年における文書の電子化と大容量化，および各種文献データベースの大規模化にともない，様々な分野において，文書データを多角的な観点から分析し，有用な知識や情報を取り出すための技術であるテキストマイニングの重要性が増してきている．その 1 つに確率的トピックモデル（probabilistic topic model，以下，トピックモデル）があり，文書を複数のトピックの混合分布として，各トピックを単語の分布として表現することによって，意味や概念に基づいた索引付けや情報検索を実現することができる．トピックモデルの先駆的な研究に，Hofmann による確率的潜在意味インデクシング法（probabilistic latent semantic indexing: pLSI）[5] がある．文書と単語はトピックのもとで条件付独立であると仮定して，各文書はトピックの混合分布，各トピックは単語の多項分布すなわちユニグラム言語モデルで表現される．トピックは観測されないため，未知パラメータを推定する必要がある．Hofmann の研究では EM アルゴリズム（expectation-maximization algorithm）の変形が用いられた．

pLSI は，基本的に，与えられた文書集合に対してトピックモデルを推定することを目的としている．このため，原理的には新たな文書に対して適切なモデルを与えるものではない．これに対して，Blei らの潜在的ディリクレ配分法（latent Dirichlet allocation: LDA）[2] では，各文書に関するトピックの多項分布にディリクレ事前分布を導入して pLSI を拡張し，新たな文書に対しても柔軟で頑健なモデリングを可能にしている．pLSI の場合と同様，LDA でもモデルパラメータを推定する必要がある．LDA モデルを推定する代表的な手法に Collapsed Gibbs Sampling 法[3),⋆1] と変分ベイズ法（variational Bayesian inference: VB）[2] があるが，最近になって，変分ベイズ法を拡張した Collapsed 変分ベイズ法（collapsed variational Bayesian inference: CVB）が提案された[4]．Collapsed 変分ベイズ法とは，変分ベイズ法よりもパラメータの独立性を緩め，潜在変数によってパラメータの依存性をモデル化することで推定精度を高めた変分アルゴリズムである．

従来研究においては，テストセットの対数尤度という観点では pLSI モデルよりも LDA モデルが性能的に優位にあり[2]，LDA モデルの推定方法としては，十分な繰返し数のもとでは変分ベイズ法や Collapsed 変分ベイズ法よりも Collapsed Gibbs Sampling 法の方が性

†1 神戸大学大学院工学研究科情報知能学専攻
Graduate School of Engineering, Kobe University

⋆1 単に Gibbs Sampling 法と呼ばれることもある[3]．

© 2009 Information Processing Society of Japan

能的に優位にある[4] とされている．しかし，単語の関連度を発見するようなタスクでの評価は十分に行われていない．本研究では Collapsed 変分ベイズ法により LDA モデルの推定を行うことで，Collapsed Gibbs Sampling 法を利用し LDA モデルを推定した場合や，従来手法である pLSI を用いる場合と比較して，上述のようなタスクに基づいた有効性を様々な観点から評価する．

以上に述べた比較評価のための実験において，本論文では医学生物学文献を用いた．当該文献では専門用語の表記が多様で，類義語が多く用いられるため，意味や概念に基づいた処理を実現するトピックモデルが特に有用であると考えられる．また，自然言語のみで知見を記述することの多い医学生物学分野において，テキストマイニングを必要とする傾向は著しい．たとえば，数百，あるいは数千種類もの遺伝子を対象にするような実験を行う場合，膨大な数の遺伝子の中から新たな発見が期待できる遺伝子の組合せを選定したり，実験によって導出された結果を解釈したりするには，関連論文の検索と精査に並々ならぬ時間と労力を費やす必要がある．また，専門分野が細分化していることによって，各分野間で情報を共有することが困難となるため，分野の垣根を超えた知見に気づかないこともありうる．たとえ同一の分野であっても，関連する論文や資料などの総量が個人が把握できる限界をはるかに超えてしまっているせいで，知見を見逃してしまっているという可能性もある．以上の点から，異なる文献に蓄積されている知見を組み合わせることによる仮説生成の可能性が論じられるようになってきた[1]．しかしながら，医学生物学文献における生物学的エンティティの相互関係予測問題の解決手段としての LDA の有効性に関しては，我々の知る限り，これまで検討されてこなかった．したがって，本論文では医学生物学文献を対象とし，生物学的エンティティ，特にタンパク質間の相互関係予測に焦点を当て，その解決手段としての LDA の有効性を分析する．

2. 関 連 研 究

本章では，まず LDA の形式化と Collapsed Gibbs Sampling 法を用いた LDA の推定方法についてその概要を示す．次に，タンパク質相互関係予測についての関連研究と本研究の位置付けを示す．

2.1 LDA の形式化

LDA のグラフィカルモデル表現を**図 1** に示す．なお，グラフィカルモデルとは，確率変数またはパラメータを頂点とし，それらの依存関係を有向グラフで表現したものである．網掛けの頂点は観測変数，それ以外の頂点は潜在変数または未知パラメータを示す．矩形部分は，その隅に示された回数だけサンプリングが繰り返されることを表す．ただし，図中の K はトピック数，D は文書数，N_j は文書 j の延べ語数を示す．LDA モデルによる文書生成過程を書き下すと，以下のようになる．

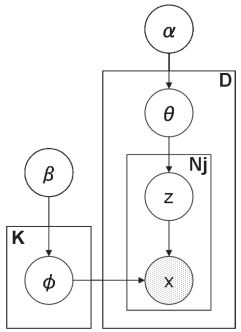

図 1　LDA のグラフィカルモデル
Fig. 1　The graphical model of LDA.

(1) 超パラメータ α を与えたディリクレ分布から各文書 j について θ_j をサンプリングする．
(2) 超パラメータ β を与えたディリクレ分布から各トピック k について ϕ_k をサンプリングする．
(3) 文書 j 内の N_j 個の語 x_i それぞれに対して
　(a) パラメータ θ_j を与えた多項分布からトピック z_i をサンプリングする．
　(b) パラメータ ϕ_{z_i} を与えた多項分布から語 x_i をサンプリングする．

LDA の全パラメータと確率変数の同時分布は，次のようになる．

$$p(\mathbf{x},\mathbf{z},\theta,\phi|\alpha,\beta)=\prod_{j=1}^{D}\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}\theta_{jk}^{\alpha-1+n_{jk\cdot}}\times\prod_{k=1}^{K}\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\prod_{w=1}^{W}\phi_{kw}^{\beta-1+n_{\cdot kw}} \quad (1)$$

このとき，W は総語彙数，n_{jkw} は文書 j 中に存在するトピック k に属する単語 w の数であり，ドット $(\cdot)$ は一致するインデックスが総計されることを意味する．つまり $n_{\cdot kw}=\sum_j n_{jkw}$，$n_{jk\cdot}=\sum_w n_{jkw}$ である．

実際の単語 $\mathbf{x}=\{x_{ij}\}$ が与えられるとき，潜在的トピックインデックス $\mathbf{z}=\{z_{ij}\}$，混合比 $\theta=\{\theta_j\}$ とトピックパラメータ $\phi=\{\phi_k\}$ 上の事後分布を計算するようなベイズ推定を

© 2009 Information Processing Society of Japan

考える．その実現方法が次に述べる Collapsed Gibbs Sampling 法であり，また，3.1 節で述べる Collapsed 変分ベイズ法である．

2.2 LDA のための Collapsed Gibbs Sampling 法

潜在変数 $\mathbf{z}$ とパラメータ θ，ϕ をサンプリングする Gibbs Sampling 法は，パラメータと潜在変数の間に強い依存性があるせいで収束が遅くなってしまう．θ と ϕ を周辺化することで，その点を改善したのが Collapsed Gibbs Sampling 法である[3]．$\mathbf{x}$ と $\mathbf{z}$ の周辺分布は，

$$p(\mathbf{z},\mathbf{x}|\alpha,\beta) = \prod_j \frac{\Gamma(K\alpha)}{\Gamma(K\alpha+n_{j\cdot\cdot})}\prod_k \frac{\Gamma(\alpha+n_{jk\cdot})}{\Gamma(\alpha)} \times \prod_k \frac{\Gamma(W\beta)}{\Gamma(W\beta+n_{\cdot k\cdot})}\prod_w \frac{\Gamma(\beta+n_{\cdot kw})}{\Gamma(\beta)} \tag{2}$$

となる．変数 z_{ij} を除くすべての変数の現在の状態が与えられたとき，z_{ij} の条件付き確率は

$$p(z_{ij}=k|\mathbf{z}^{\neg ij},\mathbf{x},\alpha,\beta) = \frac{(\alpha+n_{jk\cdot}^{\neg ij})(\beta+n_{\cdot kx_{ij}}^{\neg ij})(W\beta+n_{\cdot k\cdot}^{\neg ij})^{-1}}{\sum_{k'=1}^{K}(\alpha+n_{jk'\cdot}^{\neg ij})(\beta+n_{\cdot k'x_{ij}}^{\neg ij})(W\beta+n_{\cdot k'\cdot}^{\neg ij})^{-1}} \tag{3}$$

となる．このとき，添字 $\neg ij$ が付いたものは x_{ij} や z_{ij} を除いた変数や頻度に相当する．z_{ij} の条件付き分布は確率計算が容易な多項分布であり，プログラミングと計算にかかるコストは変分ベイズ法と比べて小さくなる．

2.3 文献によるタンパク質相互関係予測

近年，医学生物学文献からの潜在的知識発見や仮説生成の研究がさかんになってきている．これは各種文献の電子化・大容量化や，文献総数の指数関数的な増加にともない，医学生物学文献からの遺伝子機能情報の抽出やタンパク質間相互関係の抽出，疾患情報の抽出などに関する自然言語処理やテキストマイニングの必要性が高まっているためである．本研究では LDA を基にして，文献で扱われている複数のタンパク質に何らかの潜在的な関係があるかどうかを予測する．

医学生物学文献からのテキストマイニング[6),7)] では，(1) 手作業または自動的に作成されたテンプレートに基づく手法や，(2) 人間が日常的に使っている自然言語をコンピュータに処理させる自然言語処理に基づく手法，そして (3) 数理的・統計的なモデルやアルゴリズムを利用する手法などがある．自然言語処理は，エンティティ間の関係を抽出できるような構造に文書を分解するために，大量の構文解析を実行する．一方，統計的手法は，頻繁に共起するエンティティ対を発見することで互いの関係を推定する．本研究は，上記の (3) に位置づけられ，統計的手法に基づいたアプローチを用いるものであるが，これまでの当該分野の研究には見られなかった LDA の適用を試みる．

3. LDA に基づく生物学的エンティティのリンク予測

3.1 LDA に対する Collapsed 変分ベイズ法

本節では，Teh らによって提案された Collapsed 変分ベイズ法[4] について概要を示す．

Collapsed 変分ベイズ（CVB）推定とは，変分ベイズ法の考え方を基に，パラメータ ϕ，θ を周辺化するなど Collapsed Gibbs Sampling 法で行われたような手法によって推定精度を高めたアルゴリズムである．このアルゴリズムは Collapsed Gibbs Sampling 法とは異なり，潜在変数でパラメータの依存性をモデル化している．

Collapsed 変分ベイズ法では，潜在変数 $\mathbf{z}$ が互いに独立であることだけを仮定したうえで，事後分布を

$$q(\mathbf{z},\theta,\phi) = q(\theta,\phi|\mathbf{z})\prod_{ij} q(z_{ij}|\gamma_{ij}) \tag{4}$$

として近似する．このとき，$q(z_{ij}|\gamma_{ij})$ はパラメータ γ_{ij} の下での多項分布である．変分自由エネルギーは

$$\begin{aligned} F(q(\mathbf{z})q(\theta,\phi|\mathbf{z})) &= E_{q(\mathbf{z})q(\theta,\phi|\mathbf{z})}[-\log p(\mathbf{x},\mathbf{z},\theta,\phi|\alpha,\beta)] - H(q(\mathbf{z})q(\theta,\phi|\mathbf{z})) \\ &= E_{q(\mathbf{z})}[E_{q(\theta,\phi|\mathbf{z})}[-\log p(\mathbf{x},\mathbf{z},\theta,\phi|\alpha,\beta)] - H(q(\theta,\phi|\mathbf{z}))] - H(q(\mathbf{z})) \end{aligned} \tag{5}$$

となる．$q(\theta,\phi|\mathbf{z})$ に関して変分自由エネルギーを最小化する．$q(\theta,\phi|\mathbf{z})$ の分布の型に仮定をおいていないので，最小値は真の事後分布 $q(\theta,\phi|\mathbf{z}) = p(\theta,\phi|\mathbf{x},\mathbf{z},\alpha,\beta)$ によって導出する．そして，変分自由エネルギーについては以下のように単純化できる．

$$\min_{q(\theta,\phi|\mathbf{z})} F(q(\mathbf{z})q(\theta,\phi|\mathbf{z})) = E_{q(\mathbf{z})}[-\log p(\mathbf{x},\mathbf{z}|\alpha,\beta)] - H(q(\mathbf{z})) \tag{6}$$

γ_{ij} に関して式 (6) を最小化すると，次式を得る．

$$\gamma_{ijk} = q(z_{ij}=k) = \frac{\exp(E_{q(\mathbf{z}^{\neg ij})}[\log p(\mathbf{x},\mathbf{z}^{\neg ij},z_{ij}=k|\alpha,\beta)])}{\sum_{k'=1}^{K}\exp(E_{q(\mathbf{z}^{\neg ij})}[\log p(\mathbf{x},\mathbf{z}^{\neg ij},z_{ij}=k'|\alpha,\beta)])} \tag{7}$$

式 (2) を代入し，正の実数 η と正の整数 n に関して $\log\frac{\Gamma(\eta+n)}{\Gamma(\eta)} = \sum_{l=0}^{n-1}\log(\eta+l)$ に展開できることを利用し，さらに分子と分母を整理すると，次式を得る．

© 2009 Information Processing Society of Japan

$$\gamma_{ijk} = \frac{\exp(E_{q(\mathbf{z}^{\neg ij})}[\log(\alpha + n_{jk\cdot}^{\neg ij}) + \log(\beta + n_{\cdot kx_{ij}}^{\neg ij}) - \log(W\beta + n_{\cdot k\cdot}^{\neg ij})])}{\sum_{k'=1}^{K} \exp(E_{q(\mathbf{z}^{\neg ij})}[\log(\alpha + n_{jk'\cdot}^{\neg ij}) + \log(\beta + n_{\cdot k'x_{ij}}^{\neg ij}) - \log(W\beta + n_{\cdot k'\cdot}^{\neg ij})])} \tag{8}$$

なお式 (8) の期待値項の計算については，厳密に処理しようとすると計算コストがかかりすぎるので，ガウス近似を利用して簡略化する．この手法は計算コストは最小限で済むうえに，十分な近似が得られるので非常に有効である．

$n_{jk\cdot}^{\neg ij}$ の平均と分散はそれぞれ，

$$E_q[n_{jk\cdot}^{\neg ij}] = \sum_{i' \neq i} \gamma_{i'jk}, \quad Var_q[n_{jk\cdot}^{\neg ij}] = \sum_{i' \neq i} \gamma_{i'jk}(1 - \gamma_{i'jk}) \tag{9}$$

によって求めることができる．このとき，二階のテイラー展開を用いて，ガウス近似における期待値を以下のように近似する．

$$E_q[\log(\alpha + n_{jk\cdot}^{\neg ij})] \approx \log(\alpha + E_q[n_{jk\cdot}^{\neg ij}]) - \frac{Var_q(n_{jk\cdot}^{\neg ij})}{2(\alpha + E_q[n_{jk\cdot}^{\neg ij}])^2} \tag{10}$$

この近似を式 (8) の各期待値項に適用して，以下の Collapsed 変分ベイズの式を導出する．

$$\gamma_{ijk} \propto (\alpha + E_q[n_{jk\cdot}^{\neg ij}])(\beta + E_q[n_{\cdot kx_{ij}}^{\neg ij}])(W\beta + E_q[n_{\cdot k\cdot}^{\neg ij}])^{-1} \exp\left(-\frac{Var_q(n_{jk\cdot}^{\neg ij})}{2(\alpha + E_q[n_{jk\cdot}^{\neg ij}])^2} - \frac{Var_q(n_{\cdot kx_{ij}}^{\neg ij})}{2(\beta + E_q[n_{\cdot kx_{ij}}^{\neg ij}])^2} + \frac{Var_q(n_{\cdot k\cdot}^{\neg ij})}{2(W\beta + E_q[n_{\cdot k\cdot}^{\neg ij}])^2}\right) \tag{11}$$

3.2　エンティティ間類似度

エンティティ間に関連があるかどうかを評価するために，エンティティ対を用意する．本研究においては，同一文書中に併記されているエンティティの組合せを「正しい」エンティティ対として扱う．

LDA などのトピックモデルを利用すれば，あるエンティティ対が将来において，文書中に現れる尤度を計算することは，たとえそのペアがそれまでどの文書にも存在しなかったとしても，可能である．ただし，個々のエンティティはすでに出現しているものとする．

LDA に基づいた，2 つのエンティティ間の類似度[8] は，

$$Sim1(e_i, e_j) = p(e_i|e_j)/2 + p(e_j|e_i)/2 \tag{12}$$

を用いることで測定が可能である．なお，$p(e_i|e_j)$ は

$$p(e_i|e_j) = \sum_k p(e_i|k)p(k|e_j) \tag{13}$$

を計算することで求める．

この類似度計算法は，従来研究[8),9)] において使用されたものだが，この方法の場合，類似度が少数の頻出するエンティティに依存してしまう，という問題が考えられる．たとえば，片方の条件付き確率が極端に大きな値になったとき，もう片方の確率がほとんど 0 であった場合でも，2 つの値の平均をとる計算手法では，導出される結果が十分大きな値になってしまう．そこで本論文では，より正確な類似度を導出するため，式 (12) を以下のように改良する．

$$Sim2(e_i, e_j) = p(e_i|e_j) \times p(e_j|e_i) \tag{14}$$

2 つの値の積をとるこの方法であれば，上記のような場合でも計算結果は 0 に近い値になる．

4.　データセット

本章では，実験に使用した GENIA コレクション，そして TREC コレクションと GENIA タガーについて詳細を説明する．GENIA コレクションと TREC コレクションはいずれも MEDLINE のサブセットである．MEDLINE とは米国医学図書館（National Library of Medicine）が構築した医学・生物学文献データベースで，米国をはじめ，他の 70 カ国で出版された，3,800 誌を超える最新の生物医学系ジャーナルからの引用文や要約が収められている．

4.1　GENIA コレクション

GENIA コレクション[⋆1]は，XML 形式で記述された MEDLINE のサブセットであり，手作業で DNA やタンパク質などのエンティティにタグ付けがなされている．InQuery システム[10] で使用された 418 種類のストップワードを除去し，また，10 件未満の文書にしか出現しなかったエンティティや一般語も除去した．このデータセットの概要を**表 1** に示す．当該データセットは，5.1 節で述べるとおり，テストセットの対数尤度に関する予備実験で使用した．

4.2　TREC コレクションと GENIA タガー

本論文で用いる TREC コレクションは，2004 年から 2005 年の TREC Genomics

⋆1 http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi

© 2009 Information Processing Society of Japan

表 1　GENIA コレクション
Table 1　GENIA collection.

記号	名称	数
D	文書数	2,000
W	一般語の語彙数	1,959
E	エンティティの語彙数	229
W_{freq}	一般語の総頻度	107,532
E_{freq}	エンティティの総頻度	9,377

表 2　TREC コレクション
Table 2　TREC collection.

記号	名称	数	
		2002 年分	2003 年分
D	文書数	33,000	31,000
W	一般語の語彙数	16,879	183,710
E	エンティティの語彙数	1,897	58,430
W_{freq}	一般語の総頻度	3,501,405	3,863,255
E_{freq}	エンティティの総頻度	48,457	171,056

Track[⋆1]で使用されたデータであり，GENIA コレクションとは異なる XML 形式で記述されている．本論文では，タイトルと要旨の部分を 5.2 節における実験に用いた．なお，出版年が 2002 年である文書の一部を訓練データとして，2003 年である文書の一部をテストデータとして使用した．双方のデータに対して，InQuery システムで使用されたストップワード 418 種類の除去を行った．そして，訓練データにおいて，10 件未満の文書にしか出現しなかったエンティティや一般語は除去している．本研究で使用したデータセットの詳細を，**表 2** に示す．ただし，エンティティは以下で述べる GENIA タガーに基づいて数えあげたものである．

TREC コレクションは GENIA コレクション以上に膨大な文献量を有しており，エンティティにタグ付けなどはされていない．そのため，TREC コレクションから抽出した訓練データとテストデータに対して，GENIA タガー[⋆2]と呼ばれる解析ツールを使用して自動で解析を行った．なお，このツールによるエンティティのタグ付けの精度は 70%程度である．

⋆1 http://ir.ohsu.edu/genomics/
⋆2 http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Tagger

5. 評 価 実 験

本章では，提案手法と既存手法の差異を明らかにするための実験内容，およびその結果について記述する．

5.1　テストセットの対数尤度

本節では，各推定モデルに対するテストセットの対数尤度[4] に関する実験の概要について述べる．推定されたモデルのテストセットに対する単語あたりの対数尤度は値が大きいほど，より特定性の高いモデルであることを示し，一般的により良いモデルであるとされる．また，これは言語モデルの評価で広く用いられるパープレキシティ[11] の対数と負の比例関係にある．

本実験では，Collapsed Gibbs Sampling 法と Collapsed 変分ベイズ法でそれぞれ推定した 2 種類のモデルを用いて，テストセットの対数尤度を計算した．対数尤度の導出方法については 5.1.2 項で述べる．なお，本実験には GENIA コレクションと TREC コレクションの 2 種類のデータセットを使用した．それぞれのコレクション（TREC コレクションの場合は表 2 に示した 2002 年分のデータ）の各文書における単語集合をランダムに分割して，90%を訓練セットとし，残り 10%をテストセットとした．ここでいう訓練セットとテストセットは，5.2 節で用いる訓練データ，テストデータとは異なることに注意する必要がある．

5.1.1　パラメータ設定

本実験の対数尤度の導出にあたって，LDA のトピック数を $K = 10$ と設定した．また，ディリクレ事前分布の超パラメータは，GENIA コレクションでは $\alpha = 0.1$，$\beta = 0.1$ [4] に設定し，TREC コレクションでは $\alpha = 50/K$，$\beta = 0.1$ [3] に設定した．

5.1.2　対数尤度の導出

Collapsed Gibbs Sampling 法の推定モデルに対する対数尤度は，事後確率分布から S 個のサンプルが与えられたとき，

$$p(\mathbf{x^{test}}) = \prod_{ij} \sum_{k} \frac{1}{|S|} \sum_{s=1}^{S} \theta^s_{jk} \phi^s_{kx^{test}_{ij}} \tag{15}$$

から導出することができる．なお，θ_{jk} と ϕ_{kw} はそれぞれ

$$\theta^s_{jk} = \frac{\alpha + n^s_{jk\cdot}}{K\alpha + n^s_{j\cdot\cdot}} \quad \phi^s_{kw} = \frac{\beta + n^s_{\cdot kw}}{W\beta + n^s_{\cdot k\cdot}} \tag{16}$$

を計算することで求めた．

© 2009 Information Processing Society of Japan

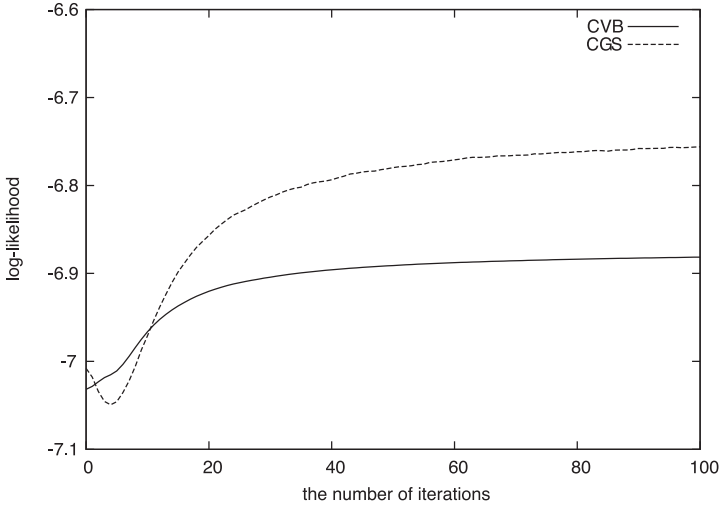

図 2　一単語あたりのテストセット対数尤度（GENIA コレクション）
Fig. 2　Per-word test-set log-likelihood over GENIA collection.

Collapsed 変分ベイズ法の推定モデルに対する対数尤度は,

$$p(\mathbf{x}^{\mathbf{test}}) = \prod_{ij} \sum_{k} \bar{\theta}_{jk} \bar{\phi}_{k x_{ij}^{test}} \tag{17}$$

から導出することができる．なお，$\bar{\theta}^s_{jk}$ と $\bar{\phi}^s_{kw}$ はそれぞれ

$$\bar{\theta}_{jk} = \frac{\alpha + E_q[n_{jk\cdot}]}{K\alpha + E_q[n_{j\cdot\cdot}]} \quad \bar{\phi}_{kw} = \frac{\beta + E_q[n_{\cdot kw}]}{W\beta + E_q[n_{\cdot k\cdot}]} \tag{18}$$

を計算することで求めた．

GENIA コレクションを用いた結果を**図 2** に示す．なお GENIA コレクションの結果は，初期値をランダムに設定して 50 回（$S = 50$）実行した平均をとっている．それぞれ 50 回実行した収束値のヒストグラムを**図 3** に示す．また，TREC コレクションを用いた結果を**図 4** に示す．

図 2 および図 4 における対数尤度のグラフでは，横軸が繰返し回数で，縦軸が 1 単語あたりの対数尤度である．図 4 において算出された対数尤度を観察すると，十数回から 20 回ぐらいまでは Collapsed 変分ベイズ法（CVB）の方が優勢だが，それ以上になると Collapsed Gibbs Sampling 法（CGS）の方が良い結果になっている．

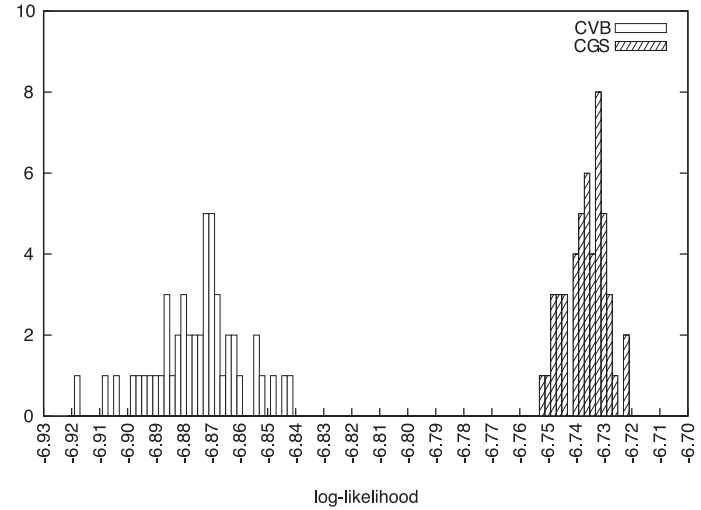

図 3　一単語あたりのテストセット対数尤度のヒストグラム（GENIA コレクション）
Fig. 3　Histogram of per-word test-set log-likelihood over GENIA collection.

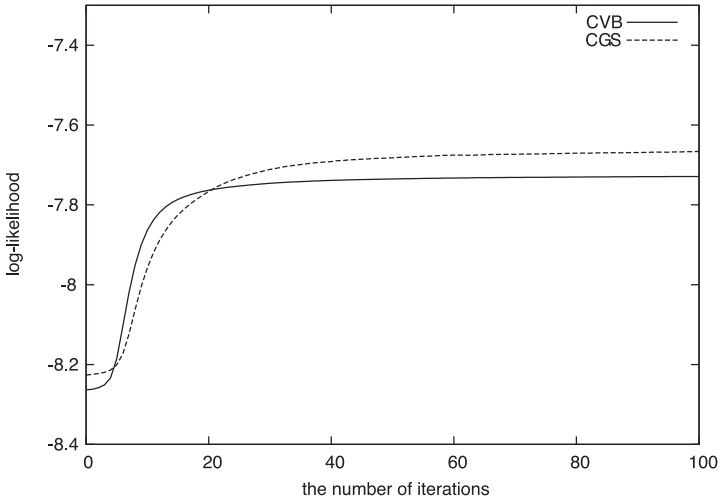

図 4　一単語あたりのテストセット対数尤度（TREC コレクション）
Fig. 4　Per-word test-set log-likelihood over GENIA collection.

© 2009 Information Processing Society of Japan

5.2　エンティティ・リンク予測

本実験では，Collapsed Gibbs Sampling 法による LDA と Collapsed 変分ベイズ法による LDA，pLSI の 3 種類を用いてエンティティ対の予測を行い，それぞれの結果を評価した．なお，pLSI の推定には EM アルゴリズムを使用した．また，3.2 節で説明した 2 種類の類似度計算手法をそれぞれ適用することで，改良した計算法で実際に精度が向上しているかどうかを確認した．なお，本実験におけるデータセットは TREC コレクションの 2002 年のデータの一部を訓練データに，2003 年のデータの一部をテストデータに用いた．これらはすでに表 2 に示したものである．本論文では特にタンパク質名のみに着目したうえで，以下で述べる実験を行った．

5.2.1　パラメータ設定

本実験のエンティティ対予測では，LDA のトピック数を $K = 10$，50，100，300 とそれぞれ設定して評価した．また，ディリクレ事前分布の超パラメータを $\alpha = 50/K$，$\beta = 0.1$ に設定し，固定した．

5.2.2　評価データの作成

本実験に際しては，エンティティ対のセットを 2 組生成した．1 つ目のセット「正解ペア」データは，訓練データでは確認されなかったが，テストデータでは確認されたエンティティ対のデータセットである．ただし，訓練データには出現しないエンティティを含んだエンティティ対は除外した．もう 1 つのセット「不正解ペア」データとは，訓練データとテストデータ双方で一度も確認されていないエンティティ対のデータセットである．本実験では，「正解ペア」データの 2 つ目のエンティティを別のエンティティへとランダムに置換し，上に述べた条件を満たすものを選別することで作成している．なお，2 つのセットに含まれるエンティティ対の数は同じで，$M = 15494$ である．

5.2.3　タスクに基づく評価

本実験では，分類精度とランキング精度に着目して 2 種類の評価尺度を利用して評価を行った．

分類精度（accuracy）の導出においては，M 個の正解ペアと M 個の不正解ペアのエンティティ間類似度をそれぞれ計算し，導出された類似度で降順に並べた．このとき，類似度が高い上位 M 個のエンティティ対を正（positive）と仮定し，それ以下の M 個のエンティティ対を負（negative）と仮定した．このとき，上位に存在する正解ペアの割合と，下位に存在する不正解ペアの割合を求めることで，分類の成否の信頼性を各パラメータごとに導出した．

表 3　分類精度
Table 3　Classification accuracy.

	CVB-LDA	CGS-LDA	pLSI
K=10	0.6310	0.6318	0.6075
K=50	0.6434	0.5359	0.5829
K=100	0.6383	0.5669	0.5648
K=300	0.6317	0.5504	0.5293

表 4　ランキング精度
Table 4　Average precision.

	CVB-LDA	CGS-LDA	pLSI
K=10	0.6651	0.6745	0.6347
K=50	0.6895	0.6574	0.5977
K=100	0.6904	0.6443	0.5606
K=300	0.6877	0.6262	0.5308

その結果，Collapsed 変分ベイズ法による LDA（CVB-LDA）の分類精度は $K = 50$ のときが最良で，それ以降はトピック数が多くなるほど分類精度は低下していることが分かった．また，Collapsed 変分ベイズ法による LDA を分析手法とした結果と，Collapsed Gibbs Sampling 法による LDA（CGS-LDA）や pLSI を分析手法とした結果を比較すると，明らかに Collapsed 変分ベイズ法による LDA の方が良い結果を示した．初期値をランダムに設定して 50 回の実験を行ったときの，それぞれの分類精度の期待値を**表 3** に示す．ただし，pLSI については結果にばらつきが生じないために 1 回の実験の分類精度を示している．

このとき，いずれのモデルや推定手法でも，トピック数が比較的少ない方が良好な結果を残した理由としては，トピック数が大きい場合にはトピックモデルが特定的になるために対数尤度は高くなるが，その反面，エンティティリンクの予測能力としては柔軟性を欠くことが理由として考えられる．

また，もう 1 つの評価指標として情報検索の評価に広く用いられる平均精度（average precision）[12] を使用した．平均精度は，正解エンティティ対のランクごとに精度（precision）を求め，正解エンティティ対データにわたって平均をとることによって求めた．本論文では分類精度に対して，上述の平均精度をランキング精度と呼ぶことにする．分類精度の場合と同じく，50 回の実験による平均精度の期待値を**表 4** に示す．Collapsed 変分ベイズ法によるランキング精度は $K = 100$ のときが最良であった．

© 2009 Information Processing Society of Japan

表 5　比較結果（Collapsed 変分ベイズ LDA の場合）
Table 5　The comparison result (the case of CVB-LDA).

		K=10	K=50	K=100	K=300
Classification	$Sim1$	0.6310	0.6434	0.6383	0.6317
Accuracy	$Sim2$	0.6351*	0.6467*	0.6398*	0.6305*
Average	$Sim1$	0.6651	0.6895	0.6905	0.6890
Precision	$Sim2$	0.6719*	0.6947*	0.6940*	0.6892

$Sim2$ に関する $Sim1$ を基準とした改善について有意差が認められた場合に $*$ を付した.

表 3 と表 4 より，Collapsed 変分ベイズ LDA（CVB-LDA）は他の 2 手法よりも高精度であることは明らかである．また，分類精度とランキング精度の両方において，Collapsed 変分ベイズ法によって推定された LDA によって得られた改善は，Collapsed Gibbs Sampling 法で推定された LDA の場合と比較して統計的に有意であることが確認できた．このとき，実験において最良の評価値であったトピック数を条件として両手法を比較した．つまり，分類精度の比較では，$K = 50$ のときの Collapsed 変分ベイズ法を $K = 10$ のときの Collapsed Gibbs Sampling 法に対して比較し，ランキング精度の比較では，$K = 100$ のときの Collapsed 変分ベイズ法を $K = 10$ のときの Collapsed Gibbs Sampling 法に対して比較した．また，検定には Wilcoxon 符号付順位検定（両側）を用いて，有意水準は 5%とした．また，Collapsed Gibbs Sampling LDA（CGS-LDA）は，分類精度とランキング精度のいずれの観点からも pLSI よりも精度が高いことが分かった．

5.2.4　類似度計算手法の比較

2 種類の LDA モデルを用いて式 (12) と式 (14) の 2 種類の計算手法によって各エンティティ対の類似度を計算し，それぞれの結果に対して分類精度とランキング精度を測定した．その結果，全体的に式 (14) を用いた方が良好な結果が得られた．こちらも各トピックごとに 50 回推定した LDA によって，統計的にはおおむね有意であることを確認している．一例として，Collapsed 変分ベイズ LDA で比較した際の結果を**表 5** に示す．

5.2.5　タンパク質相互関係ネットワーク

以上の実験では，LDA モデルの推定時に 2002 年の文書データである訓練データを使用することで，2003 年の文書データであるテストデータの知見をどれだけ推定できているかを確認した．

表 3 より，最良の分類精度であったトピック数を 50 に設定した Collapsed 変分ベイズ法によって推定したモデルを基に導出した，エンティティ・ネットワークの例を**図 5** に示す．辺の長さは類似度を示しており，辺が短いほど 2 つのエンティティ間の類似度が高いことを

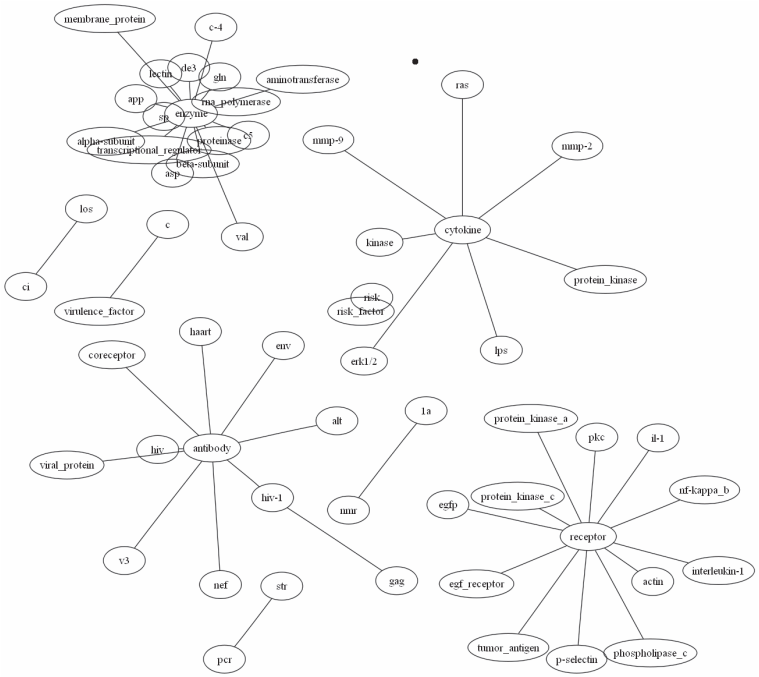

図 5　エンティティ・ネットワーク
Fig. 5　Entity-entity network.

示している．なお，図 5 は類似度で降順に並べたエンティティ対上位 50 個によって構成されたエンティティ・ネットワークである．ノードがタンパク質を示しており，各ノードをつなぐリンクが短いほど，つながれた 2 つのタンパク質の間には何らかの関連性がある可能性が高いことを示している．たとえば図の左上にある “enzyme（酵素）” を中心としたノード群は，同じく酵素の一種であるか，もしくは酵素と何らかの関わりを持つタンパク質である可能性が高い，と考えられる．ただし生成された仮説の正誤については専門家による検討を必要とするため，必ずしも短いリンクでつながっているノード間に有用な関連性があるわけではない．

© 2009 Information Processing Society of Japan

6. おわりに

本論文では，Collapsed 変分ベイズ法を利用し LDA モデルを推定することで，医学生物学文献データベースから仮説を生成する手法を提案した．生物学的エンティティ，特にタンパク質に関する相互関係を予測するタスクに焦点を当てて実験を行った．一般的に利用されている推定手法である Collapsed Gibbs Sampling 法で推定した LDA モデルや，従来手法で利用されていた pLSI と比較することで，当該タスクによる評価すなわち分類精度とランキング精度の観点から見れば，Collapsed 変分ベイズ法による LDA の方が他の 2 手法よりも良好な結果を示すことが確認できた．また，従来研究で利用されていた手法を改良した類似度計算法が改良前より良好な結果を導出できることも確認した．

テストセットの対数尤度の観点では Collapsed Gibbs Sampling 法に劣る Collapsed 変分ベイズ法が，今回のように単語の関連度を推定する実験では逆に優位となったが，その理由を理論的に解明することは今後の課題である．

それ以外にも，自然言語処理など，単語共起に基づく統計的手法とは異なる方法と組み合わせることも今後の研究の方向性として考えられる．

謝辞　本研究の一部は，科学研究費補助金特定領域研究「情報爆発 IT 基盤」（19024055），基盤研究（B）（20300038）の援助による．

参考文献

1) 小池麻子：テキストマイニングによる潜在的知識の発見支援，情報処理，Vol.48, No.8, pp.824–829 (2007).
2) Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
3) Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *Proc. National Academy of Sciences of the United States of America*, 101, pp.5228–5235 (2004).
4) Teh, Y.W., Newman, D. and Welling, M.: A collapsed variational bayesian inference algorithm for latent dirichlet allocation, *Advances in Neural Information Processing Systems*, Vol.16 (2007).
5) Hofmann, T.: Probabilistic latent semantic indexing, *Proc. 22nd International Conference on Research and Development in Information Retrieval*, Berkeley, California, USA, pp.50–57 (1999).
6) Cohen, A.M. and Hersh, W.R.: A Survey of Current Work in Biomedical Text Mining, *Briefings in Bioinformatics*, Vol.6, No.1, pp.57–71 (2005).
7) Zhou, D. and He, Y.: Extracting Interactions between Proteins from the Literature, *Journal of Biomedical Informatics*, No.41, pp.393–407 (2008).
8) Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M.: Statistical entity-topic models, *Proc. 12th ACM SIGKDD international conference on Knowledge discoverry and data mining*, Philadelphia, Pennsylvania, pp.680–686 (2006).
9) Steyvers, M. and Griffiths, T.: *Handbook of Latent Semantic Analysis*, *chapter 21: Probabilistic Topic Models*, Lawrence Erbaum Associates, Mahwah, New Jersey, London (2007).
10) Callan, J.P., Croft, W.B. and Harding, S.M.: The INQUERY Retrieval System, *Proc. 3rd International Conference on Database and Expert Systems Applications*, Valencia, Spain, pp.78–83 (1992).
11) Rabiner, L. and Juang, B.-H.: 音声認識の基礎，NTT アドバンステクノロジ株式会社 (1995).
12) Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison-Wesley (1999).
13) Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.: Indexing by Latent Semantic Analysis, *Journal of the Society for Information Science*, Vol.41, No.6, pp.391–407 (1990).
14) Homayouni, R., Heinrich, K., Wei, L. and Berry, M.W.: Gene clustering by Latent Semantic Indexing of MEDLINE Abstracts, *Bioinformatics*, Vol.21, No.1, pp.104–115 (2005).
15) Kim, H., Park, H. and Drake, B.L.: Extracting Unrecognized Gene Relationships from the Biomedical Literature via Matrix Factorizations, *BMC Bioinformatics*, Vol.8 (Suppl. 9), No.S6 (2007).

(平成 20 年 12 月 20 日受付)
(平成 21 年 4 月 6 日採録)

(担当編集委員　高久 雅生)

© 2009 Information Processing Society of Japan

麻生　竜矢

平成 20 年神戸大学工学部情報知能工学科卒業，現在，同大学大学院工学研究科情報知能学専攻博士前期課程在学中.

江口　浩二（正会員）

神戸大学大学院工学研究科情報知能学専攻准教授，国立情報学研究所客員准教授．博士（工学）．情報検索，Web 情報処理，データマイニングの研究に従事.

© 2009 Information Processing Society of Japan