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

2008－DBS－145 (8)
2008－ＦＩ－ 91 (8)
2008／6／19

Collapsed変分ベイズLDAによるタンパク質相互作用予測

麻生 竜矢† 江口 浩二†

†神戸大学大学院工学研究科情報知能学専攻
〒657–8501 神戸市灘区六甲台町 1–1
E-mail: dango-r@cs25.scitec.kobe-u.ac.jp

要旨 近年，医学生物学分野を始めとする様々な領域において，増加の一途をたどる電子化された文書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の高度化への要求が高まっている．この目的のもと，我々は確率的トピックモデルをテキストから抽出された生物学的エンティティとりわけタンパク質間の相互作用予測タスクへ適用する．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，上述のタスクに対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．本稿では，LDA の推定手法として Collapsed 変分ベイズ法とギブスサンプリング法を適用し，対数尤度，分類精度ならびにランキング精度の観点から比較を行う．特に分類精度とランキング精度については，Collapsed 変分ベイズ法によって良好な結果が得られることを確認した．

Predicting Protein-Protein Interactions using Collapsed Variational Latent Dirichlet Allocation

Tatsuya Asou† Koji Eguchi†

†Kobe University
Department of Computer Science and Systems Engineering
1-1 Rokkoudai, Nada-ku, Kobe, 657-8501, Japan
E-mail: dango-r@cs25.scitec.kobe-u.ac.jp

Abstract Recently, technologies for organizing knowledge accumulated in a growing number of digitized documents and then generating potential hypotheses have been highly requested, such as in biomedical fields. For these objectives, we investigate applying statistical topic models to predict interactions 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 and inference via Gibbs sampling to estimating the LDA model, and compared them from the viewpoints of log-likelihoods, classification accuracy and retrieval effectiveness. We demonstrate through experiments that the Collapsed Variational LDA gives better results than the other, especially in terms of classification accuracy and retrieval effectiveness.

1 はじめに

近年における文書の電子化と大容量化，及び各種文献データベースの大規模化に伴い，様々な分野においてテキストマイニングの重要性が増してきている．中でも，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，その傾向は著しいものがある．

例えば，数百，或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み合わせを選定したり，実験によって導出された結果を解釈するには，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，分野の垣根を越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遥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]．本稿では，対象とする問題として，生物学的エンティティ，特にタンパク質間の相互作用を予測する問題に焦点を当てる．

ところで一般的なテキストマイニング手法の1つとして潜在的ディリクレ配分法 (Latent Dirichlet Allocation: LDA)[2] がある．しかし，医学生物学文献に対する生物学的エンティティの相互作用予測問題に向けた解決手段としての 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．

LDA モデルを推定する方法として一般的なのが Collapsed Gibbs Sampling 法[1][3] や変分ベイズ法 [2] である．最近になって，変分ベイズ法を拡張した Collapsed 変分ベイズ法が提案された [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．

本研究では医学生物学文献データベースを対象に，Collapsed 変分ベイズ法 [4] 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と比較してその有効性を様々な観点から評価する．とりわけ，生物学的エンティティ間相互作用予測に焦点を当て，その解決手段としての有効性を分析する．

[1] 単に Gibbs Sampling 法と呼ばれることもある [3].

2 関連研究

2.1 LDA での推定アルゴリズム

トピックモデルとは「文書は，ある特徴を持った単語の分布（トピック）の混合分布から生成される」という考えに基づいたモデルである [5, 2, 6]．Blei ら [2] によって，文書のトピックを表す多項分布にディリクレ事前分布を導入する潜在的ディリクレ配分法 (Latent Dirichlet allocation: LDA) が提案された．このとき LDA モデルの推定には変分ベイズ法が使用された．LDA モデルの推定に広く用いられるもう一つの手法は Griffiths ら [3] が用いた Collapsed Gibbs Sampling 法である．また，Teh ら [4] は変分ベイズ法を拡張した Collapsed 変分ベイズ法を LDA モデルの推定に用いて，パープレキシティ[7] の観点でより良い推定結果を示した．

2.1.1 LDA での近似推定

LDA のグラフィカルモデル表現を図 1 に示す．なお，図中の D は文書数，K はトピック数，N_j は文書 j のべ語数を示す．LDA は文書をトピックの混合分布としてモデル化する．LDA の生成プロセスは以下の通りである．

(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上の完全な同時分布は，次のようになる．

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

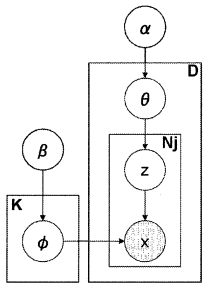

図 1: LDA のグラフィカルモデル

このとき $n_{jkw} = \sharp\{i : x_{ij} = w, z_{ij} = k\}$ で，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．

観測された単語 $\mathbf{x} = \{x_{ij}\}$ が与えられるとき，ベイズ推定のタスクとは，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ことである．

2.1.2 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の強い依存性により，緩やかに収束する．θ と ϕ を周辺化することで，それらに対処したのが Collapsed Gibbs Sampling 法である．$\mathbf{x}$ と $\mathbf{z}$ の周辺分布は，

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

となる．ある変数 z_{ij} 以外全ての現在状態が与えられたとき，z_{ij} の状態確率は

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

となる．このとき，肩付きの文字 $\neg ij$ は x_{ij} や z_{ij} を除いた変数や値に相当する．z_{ij} の状態分布は，確率計算が簡単な多項分布であり，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．

2.2 文献によるタンパク質相互作用予測

近年，医学生物学文献からの潜在的知識発見／仮説生成の研究は盛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伴い，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相互作用の抽出，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為である．本研究では LDA を元にして，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 [8] 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・統計的モデル、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．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] について説明を行う．

Collapsed Gibbs Sampling 法は Gibbs Sampling 法よりも収束が早いことが確認されている．前章の (3) 式より，z_{ij} は $n_{jk\cdot}^{\neg ij}$ と $n_{\cdot kx_{ij}}^{\neg ij}$，$n_{\cdot k\cdot}^{\neg ij}$ を通して $z^{\neg ij}$ に依存することに注意しなければならない．特に，どんな特殊な変数 $z_{i'j'}$ であっても z_{ij} の依存性はとても弱く，大きいデータセットにおいては特に弱い．その結果，Collapsed Gibbs Sampling 法の収束が早くなると予想される．しかしながら，他のマルコフ連鎖モンテカルロ (Markov Chain Monte Carlo: MCMC) サンプラーと，MCMC とは異なる変分推定について，収束性の理由を究明するのは非常に困難であり，またサンプリングノイズを除去するためには大量のサンプルが必要となる可能性がある．

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 法のように，

独立性を仮定する代わりに潜在変数でパラメータの依存性をモデル化している．

パラメータを処理する方法は二つあり，一つ目は同時分布を周辺化して前章の(2)式から始めるもので，二つ目は，その事後分布の形式においてどんな仮定もしないzとxを与えられたθとφの，事後分布をモデル化することである．ここでは，それら二つの方法が同義であることを示す．

Collapsed変分ベイズ法で行う唯一の仮定は，潜在変数zが互いに独立であるということだけであり，したがって事後分布を

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

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

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

となる．最初に$\hat{q}(\theta,\phi|\mathbf{z})$に関して変分自由エネルギーを最小化し，続けて$\hat{q}(\mathbf{z})$を最小化する．$\hat{q}(\theta,\phi|\mathbf{z})$の形式の制約はないので，最小値は真の事後分布$\hat{q}(\theta,\phi|\mathbf{z}) = p(\theta,\phi,\mathbf{x},\mathbf{z}|\alpha,\beta)$で達成される．そして，変分自由エネルギーは以下のように単純化する．

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

Collapsed変分ベイズ法はz上の事後分布を近似する前にθとφを周辺化することと同義である，と考えられる．Collapsed変分ベイズ法は変分ベイズ法よりも，変分事後分布における仮定が厳密に弱くなるので，次式を得る．

$$\widehat{F}(\hat{q}(\mathbf{z})) \leq \widetilde{F}(\tilde{q}(\mathbf{z})) \triangleq \min_{\tilde{q}(\theta)\tilde{q}(\phi)}\widetilde{F}(\tilde{q}(\mathbf{z})\tilde{q}(\theta)\tilde{q}(\phi)) \tag{7}$$

したがって，Collapsed変分ベイズ法は普通の変分ベイズ法よりも良好な近似となる．$\hat{\gamma}_{ij}$に関して(7)式を最小化するとき，次式を得る．

$$\begin{aligned}\hat{\gamma}_{ijk} &= \hat{q}(z_{ij}=k) \\ &= \frac{\exp(E_{\hat{q}(\mathbf{z}^{\neg ij})}[p(\mathbf{x},\mathbf{z}^{\neg ij},z_{ij}=k|\alpha,\beta)])}{\sum_{k'=1}^{K}\exp(E_{\hat{q}(\mathbf{z}^{\neg ij})}[p(\mathbf{x},\mathbf{z}^{\neg ij},z_{ij}=k'|\alpha,\beta)])}\end{aligned} \tag{8}$$

(2)式を代入し，正の実数ηと正の整数nに関して$\log\frac{\Gamma(\eta+n)}{\Gamma(\eta)} = \sum_{l=0}^{n-1}\log(\eta+l)$を拡張して，分子と分母の両方にある項を消去すると，次式が得られる．

$$\hat{\gamma}_{ijk} = \frac{\exp(E_{\hat{q}(\mathbf{z}^{\neg ij})}[\log(\alpha+n_{jk\cdot}^{\neg ij})+\log(\beta+n_{\cdot kx_{ij}}^{\neg ij})-\log(W\beta+n_{\cdot k\cdot}^{\neg ij})])}{\sum_{k'=1}^{K}\exp(E_{\hat{q}(\mathbf{z}^{\neg ij})}[\log(\alpha+n_{jk'\cdot}^{\neg ij})+\log(\beta+n_{\cdot k'x_{ij}}^{\neg ij})-\log(W\beta+n_{\cdot k'\cdot}^{\neg ij})])} \tag{9}$$

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為に，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み合わせを“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既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[9]は，

$$p(e_i|e_j)/2 + p(e_j|e_i)/2 \tag{10}$$

を用いることで測定が可能である．なお，$p(e_i|e_j)$は

$$p(e_i|e_j) = \sum_k p(e_i|k)p(k|e_j) \tag{11}$$

を計算することで求める．

4 データセットとGENIAタガー

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ヶ国で出版された，3,800誌を越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[2]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為されている．InQuerryシステムで使用された418個のストップワードを除去し，また，10より少ない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詳細を，表1に示す．

[2]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi

表 1: GENIA コレクション

D	文書数	2000
W	一般語の語彙数	1959
E	エンティティの語彙数	229
W_{freq}	一般語の総数	107532
E_{freq}	エンティティの総数	9377

4.2 TREC コレクションと GENIA タガー

4.2.1 TREC コレクション

本稿で用いる TREC コレクションは，2004 年から 2005 年の TREC Genomics Track[3]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．

本研究では，タイトルと本文，もしくは要旨の部分を利用した．また，PubDate タグと DataComplete タグ内の Year タグが共に 2002 年である文書の一部を 5.2 節における訓練データとして，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 InQuery システムで使用された，418 個のストップワードを除去した．そして，訓練データにおいて，10 より少ない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，エンティティは次節に述べる GENIA タガーに基づいて数え上げたものである．

4.2.2 GENIA タガー

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．その為，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[4]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．

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為の実験内容，及びその結果について記述する．

[3]http://ir.ohsu.edu/genomics/

[4]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Tagger

表 2: TREC コレクション

		2002 年分	2003 年分
D	文書数	33000	31000
W	一般語の語彙数	16879	183710
E	エンティティの語彙数	1897	58430
W_{freq}	一般語の総数	3501405	3863255
E_{freq}	エンティティの総数	48457	171056

5.1 対数尤度

本節では，各推定モデルに対する対数尤度の導出について記述する．推定されたモデルのテストデータ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，一般的によりよいモデルであるとされる．また，言語モデルの評価で広く用いられるパープレキシティ[7] の対数と負の比例関係にある．

本実験では，Collapsed Gibbs Sampling 法と Collapsed 変分ベイズ法の 2 つの推定アルゴリズムで推定したモデルの，対数尤度 [4] を計算した．対数尤度の詳細については 5.1.2 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．各コレクション（TREC コレクションは表 2 に示した 2002 年分のデータ）をランダムに分割したデータの 90 %を訓練データとし，残り 10 %をテストデータとした．ここでいう訓練データとテストデータは，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せよ．

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 $\alpha = 50/K$，$\beta = 0.1$ 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 $S = 1$ 個のサンプルが与えられたとき，

$$p(\mathbf{x}^{\mathbf{test}}) = \prod_{ij} \sum_{k} \bar{\theta}_{jk} \bar{\phi}_{k x_{ij}^{test}} \tag{12}$$

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

$$\bar{\theta}_{jk} = \frac{\alpha + E_q[n_{jk\cdot}]}{K\alpha + E_q[n_{j\cdot\cdot}]} \quad \bar{\phi}_{kw} = \frac{\beta + E_q[n_{\cdot kw}]}{W\beta + E_q[n_{\cdot k\cdot}]} \tag{13}$$

を計算することで求めた．

Collapsed 変分ベイズ法の推定モデルに対する対数尤度は，

$$p(\mathbf{x^{test}}) = \prod_{ij} \sum_{k} \frac{1}{|S|} \sum_{s=1}^{S} \bar{\theta}^s_{jk} \bar{\phi}^s_{kx^{test}_{ij}} \tag{14}$$

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

$$\bar{\theta}^s_{jk} = \frac{\alpha + n^s_{jk\cdot}}{K\alpha + n^s_{j\cdot\cdot}} \quad \bar{\phi}^s_{kw} = \frac{\beta + n^s_{\cdot kw}}{W\beta + n^s_{\cdot k\cdot}} \tag{15}$$

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 回実行した平均を取っている．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．

算出された対数尤度のグラフは，横軸が繰り返し回数で，縦軸が 1 単語あたりの対数尤度である．それぞれ算出された対数尤度を比較すると，十数回から二十回ぐらいまでは 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 (GS) の方が良い結果になった．

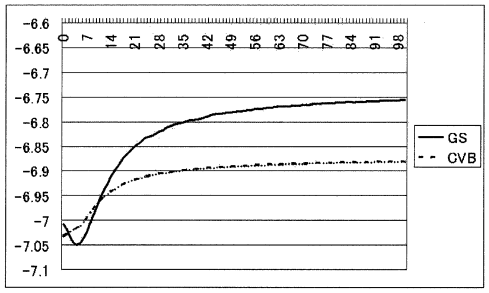

図 2: 一単語あたりのテストセット対数尤度 (GENIA コレクション)

5.2 エンティティ・リンク予測

本実験では，Collapsed Gibbs Sampling 法と Collapsed 変分ベイズ法の 2 種類のアルゴリズムで推定したモデルを用いてエンティティ対の予測を行い，それぞれ評価した．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．これらは既に表 2 に示したものである．

本稿では特にタンパク質名のみに着目して次節以降で述べる実験を行った．

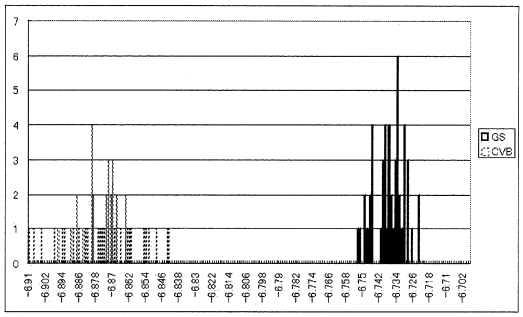

図 3: 一単語あたりのテストセット対数尤度のヒストグラム (GENIA コレクション)

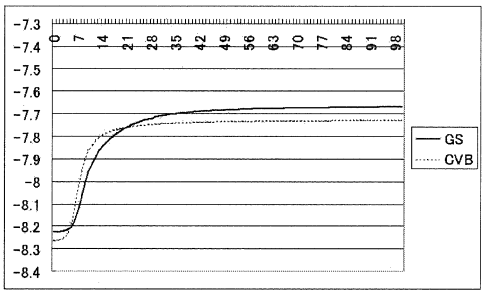

図 4: 一単語あたりのテストセット対数尤度 (TREC コレクション)

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 $K = 50$，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”データの二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

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，導出された類似度の降順に並べた．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．このときの実際の正解ペアと不正解ペアと比べた分類精度 (accuracy) を求めることで成否の信頼性を各パラメータごとに導出した．

その結果，分類精度は $K = 50$ のときが最も良く，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わかった．また，今回の提案手法である Collapsed 変分ベイズ法を分析手法とした結果と Collapsed Gibbs Sampling 法を分析手法とした結果を比較すると，明らかに Collapsed 変分ベイズ法が良い結果となった．表3にそれぞれの分類精度の結果を記す．

このとき，トピック数が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，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．

表 3: Classification Accuracy

トピック数	CollapsedGibbsSampling	CollapsedVB
$K = 50$	0.5799	0.6479
$K = 100$	0.5623	0.6426
$K = 300$	0.5477	0.6335

また，もう一つの評価指標として情報検索の評価に広く用いられる average precision[10] を使用した．average precision とは，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．本稿ではこれをランキング精度と呼ぶこともある．表4に average precision の評価結果を記す．ランキング精度では $K = 100$ のときが最良であった．

表 4: Average Precision

トピック数	CollapsedGibbsSampling	CollapsedVB
$K = 50$	0.6464	0.6808
$K = 100$	0.6472	0.6984
$K = 300$	0.6223	0.6903

5.2.4 タンパク質相互作用ネットワーク

本実験では，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に強い関係があることを示す．

なお，図5は類似度が上位 50 個のエンティティ対によって構成されたエンティティ・ネットワークである．

6 おわりに

本稿では，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とくにタンパク質に関する相互作用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 モデルと比較することで，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 変分ベイズ法の方が良好な結果を示すことが確認できた．

今後の課題としては，まず，Latent Semantic Indexing[11] などの既存手法 [12, 13] と比較することで，導出される結果にどれだけの差異や有効性の違いが見られるかを確認することが考えられる．

また，今後の研究のもう一つの方向性として，本稿で提案した医学生物学エンティティ・ネットワークの予測手法をもとに，エンティティ・ネットワークのブ

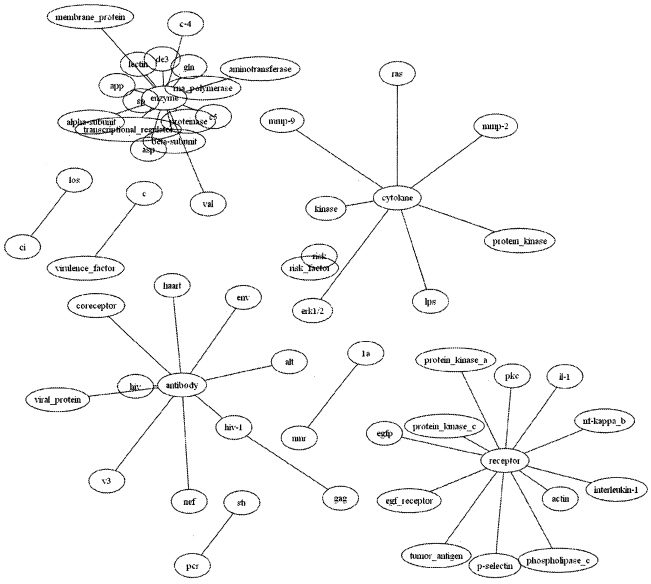

図 5: エンティティ・ネットワーク

ラウジングや関連文献の検索を可能とするシステム「BioNetPro（Biomedical entity Network analyzer based on Probability）」を開発することが挙げられる．

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, 48 巻 8 号, pp. 824–829 (2007).

[2]D.M.Blei, A.Y.Ng and M.I.Jordan, “Latent dirichlet allocation”, Journal of Machine Learning Research, 3, pp. 993–1022(2003).

[3]T.L.Griffiths and M.Steyvers, “Finding scientific topics”, Proceedings of the National Academy of Sciences of the United States of America, 101, pp. 5228–5235 (2004).

[4]Y.W.Teh,D.Newman and M.Welling, “A collapsed variational bayesian inference algorithm for latent dirichlet allocation”, In NIPS 2006:Neural Information Processing Systems Conference 2006 (2006).

[5]T.Hofmann, “Probabilistic latent semantic indexing”, In Proceeding of the 22nd International Conference on Research and Development in Information Retrieval, Berkeley, California, USA, pp. 50–57 (1999).

[6]M.Steyvers and T.Griffiths, “Handbook of Latent Semantic Analysis”, chapter 21:Probabilistic Topic Models, Lawrence Erbaum Associates, Mahwah, New Jersey, London (2007).

[7]Lawrence Rabiner, Biing-Hwang Juang, “音声認識の基礎” NTT アドバンステクノロジ株式会社 (1995).

[8]Aaron M. Cohen, and William R. Hersh, “A Survey of Current Work in Biomedical Text Mining”, Briefings in Bioinformatics, Vol.6, No.1, pp. 57–71 (2005).

[9]D.Newman, C.Chemudugunta, P.Smyth and M.Steyvers, “Statistical entity-topic models”, In Proceeding of the 12th ACM SIGKDD international conference on Knowledge discoverry and data mining, New York, NY, USA, ACM Press, pp. 680–686(2006).

[10]Ricardo Baeza-Yates, and Berthier Ribeiro-Neto, “Modern Information Retrieval”, Addison-Wesley (1999).

[11]S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, “Indexing by Latent Semantic Analysis”, Journal of the Society for Information Science, 41(6), pp. 391–407 (1990).

[12]Ramin Homayouni, Kevin Heinrich, Lai Wei, and Michael W. Berry, “Gene clustering by Latent Semantic Indexing of MEDLINE Abstracts”, Bioinformatics, Vol.21, No.1, pp. 104–115 (2005).

[13]Hyunsoo Kim, Haesun Park, and Barry L Drake, “Extracting Unrecognized Gene Relationships from the Biomedical Literature via Matrix Factorizations”, BMC Bioinformatics, Vol.8 (Suppl 9), No.S6 (2007).