Vol.2013-NL-214 No.1
2013/11/14

フレーズ生成機構を組み込んだ潜在変数を有する生成モデルによるトピック分析

濱田 龍之介 [1, a)]　船津 繁晃 [1, b)]　冨浦 洋一 [1, c)]

キーワードで論文を検索した際には，膨大な量の検索結果が得られる．その中から利用者が欲している情報を探すためにそれら全てのアブストラクトを読んで確認するのは大変な労力を要する．そこで，文書集合内のトピックを推定し，利用者に該当するトピックを指定させ，指定したトピックを含む論文に絞り込むことが考えられる．トピック分析の代表的な手法である LDA では，それぞれのトピックで高い確率で出現する語を提示することによってトピックの意味内容を把握させる．しかし，キーワード検索でヒットした論文アブストラクト集合は，内容が類似しているため，前述のような単語を提示する方法ではトピックの内容を把握するのは難しい．そこで，フレーズを生成する仕組みをモデルに組み込んで LDA を拡張することにより，複数の内容語から成るフレーズを提示することで，トピックの内容理解を容易にする手法を提案する．

Topic Analysis using the Latent Variable Model with a Mechanism for Generation Phrases

Ryunosuke Hamada[1, a)]　Toshiaki Funatsu[1, b)]
Yoichi Tomiura[1, c)]

When users perform a keyword search for scientific papers, results of the search are huge amounts. They spend a lot of energy finding the appropriate papers that satisfy their information demand. If we can estimate the topics in the abstracts in the results of the search, we can narrow the result to papers including a selected topic. In LDA (Latent Dirichlet Allocation), a typical method of topic analysis, each topic is represented by the words that appear in the topic with high probability. However, it is difficult for users to grasp the meaning of the topic with such words in the case of topic analysis for abstracts in the results of the keyword search. This is because the abstracts in the results are similar to each other. Hence, we propose a method to facilitate our understanding meaning of topics by incorporating a mechanism for generating phrases into LDA and presenting the phrases instead of words.

1. はじめに

ウェブ上の検索エンジンの普及に伴い，キーワードによる検索で求める情報に容易にアクセスできるようになった．しかし，学術論文の検索の場合，キーワードによる検索では応えることが困難な情報要求も多い．研究の新規性を確認したり，問題解決のための理論や手法を探索したりする際の情報要求は非常に限定的なものになる．そのような情報要求を検索者自身が言語化したキーワードによる検索では，求める情報を記述した論文の著者がそれと同じ用語を用いているとは限らないため，情報要求を満たす論文が漏れてしまう可能性が高い．一方，情報要求を少し一般化したキーワードによる検索を行った場合は，大量の論文が該当してしまい，しかも，その多くが当初の限定的な情報要求を満たすものではない．したがって，情報要求を満たす論文か否かを確認する作業に多大な労力を費やすことになる．そこで，この作業を容易にするために，得られた検索結果のアブストラクト集合に含まれるトピックを分析し，利用者に該当するトピックを指定させ，指定されたトピックを含むアブストラクトの論文に絞り込むことが考えられる．

トピック分析の代表的な手法である LDA[1][2] では，それぞれのトピックで高い確率で出現する単語のリストからトピックの意味内容を解釈することになる．しかし，キーワード検索でヒットした論文アブストラクト集合は，文書内の単語，文書数ともに少なく，かつ類似した専門的な内容の文書であるため，前述のような単語を提示する方法ではトピックの内容を把握するのは難しい．

そこで本研究では，トピックの内容把握を容易にするために，フレーズ（複合的な表現）を生成する仕組みをモデルに組み込んで LDA を拡張することにより，トピックの意味内容を解釈するための情報として，単語またはフレーズ（複合的表現）を提示する手法を提案する．また，前処理として，C-value[3]などを用いて単語列をフレーズの列に変換し，通常の LDA でトピック推定を行う実験を行い，前述の実験との比較考察を行う．

以下 2 節では LDA の説明及びその問題点について述べ，3 節で関連研究，4 節で提案手法についての説明，5 節でトピック分析例とその評価，6 節で本稿のまとめと今後の展望について述べる．

2. 従来手法(LDA)の問題点

LDA(Latent Dirichlet Allocation) では，各文書がトピックの混合比を持ち，各トピックが単語の出現確率分布を持つ

1 九州大学
Kyushu University
a) hamada@nlp.inf.kyushu-u.ac.jp
b) funatsu@nlp.inf.kyushu-u.ac.jp
c) tom@inf.kyushu-u.ac.jp

とする．そして，文書の生成では，文書の各単語位置に対して，文書のトピックの混合比に従ってトピックが生成され，そのトピックが持つ単語の出現確率分布に従ってその位置の単語が生成されると考える．トピック分析では，文書集合とトピック数を与え，Gibbs Sampling を利用して，各文書の各単語位置に対するトピックを割り当て，それにより，文書ごとのトピックの混合比およびトピックごとの単語の出現確率分布を推定する．

LDA でトピック分析を行う際, その対象となる文書集合の内容に大きな違いがある場合は，トピックに頻出する単語を提示するだけでそのトピックの内容を解釈することができる．以下に J-STAGE[4]から言語処理学会の学会誌「自然言語処理」より 421 件の論文アブストラクトを収集し，その集合に対して LDA によるトピック分析をトピック数 10 として行った結果の一部を示す．

表 1 の単語リストより，Topic1 が「文書検索」Topic2 が「機械翻訳」，Topic3 が「音声認識」についてのトピックだと解釈できる. 言語処理学会の学会誌の論文と言えども，その内容は様々で，このように，内容に大きな違いがある文書集合に対しては, LDA によるトピック分析が非常に有効だということが分かる．

次に,「NTCIR-1: 情報検索/用語抽出研究用テストコレクション」より用語「機械翻訳」を含む 636 件のアブストラクトを取り出し，これをクエリ「機械翻訳」で論文を検索した結果のアブストラクト集合と見なす．この文書集合に対して LDA によるトピック分析をトピック数 10 にして行った結果の一部を表 2 に示す．この単語リストを見ただけでは，それぞれのトピックの内容を解釈するのは困難である．

表 1 「自然言語処理」に関する論文集合に対してのトピック分析の結果

Topic1	Topic2	Topic3
検索	翻訳	対話
手法	言語	音声
類似	システム	言語
文書	機械	特徴
提案	日本語	認識
単語	辞書	文
情報	対訳	手話
精度	文	システム
本	英語	発話
方法	日	質問
問題	英	感情
システム	知識	テキスト
入力	訳語	ユーザ
実験	対応	表現
候補	実験	動作

表 2 「機械翻訳」に関する論文集合に対してのトピック分析の結果

Topic1	Topic2	Topic3
翻訳	翻訳	言語
システム	文	概念
機械	表現	自然
ユーザ	解析	処理
こと	日本語	意味
機能	こと	知識
文書	機械	システム
作成	構造	理解
結果	処理	ため
ため	システム	こと
情報	生成	中間
支援	日	研究
利用	構文	情報
必要	言語	関係
英	提案	文法

キーワードによる検索結果のアブストラクト集合のトピック分析において, トピックの内容解釈が困難な原因は，キーワードによる検索結果のアブストラクト集合中の文書は類似したものとなり，トピック分析では，より細かな専門的なトピックに分解されることによる. 実際の検索では，「機械翻訳」よりもさらに限定したクエリを用いると考えられ，その場合はさらに細かな専門的なトピックになる．そのような専門的な内容のトピックを表現するには，

(a) 単語だけではなくフレーズ（複合的な表現）による表示

(b) そのトピックで特徴的に出現する単語またはフレーズだけに絞った表示

(c) 表示される語の間の抄録中での依存関係の表示

が必要と考えられる. 本研究は上記の(a)(b)への対処を目的として行った．

3. 関連研究

トピックの内容を推定するために外部知識を用いるという方法もある．例えば文献[5]では，外部知識である Wikipedia を知識源とした「分野トピックモデル」なるものを提案し，文書集合のトピックを推定している．しかしながら, Wikipedia には新規に提案された手法やあまり一般的ではない用語は登録されておらず、新規手法や一般的ではない用語が頻出する文書集合に対しては，うまく機能しない可能性が高い. このような用語にも対応できるところが，今回の手法の利点の一つである．

また，トピック推定をする際に bag-of-words ではなく，順次情報によって濃縮されたコンテキストを利用することで, より洗練されたモデリングを提示しているものもある，文献[6]では, トピック推定を単語 bigram やトピック bigram

単位で行うことにより，通常の LDA よりも精度の高いトピック分析を可能にしている．今回の手法では文脈情報を用いるだけでなく，トピック理解の容易性を考えてフレーズを生成してそれを利用者に提示するということが特徴的である．

4. トピックの解釈容易性を考慮したトピック分析手法

本研究では，トピックの解釈の容易性を考慮し，複合的表現(フレーズ)を生成する機構をトピック毎の単語bigramとして導入することで LDA の拡張を行った．

LDA と同じく，文書を単語の列と捉える．文書集合を $\boldsymbol{w}$ で表し，m 番目の文書を $\boldsymbol{w}^{(m)}$ で表す．文書数を M とすると，$\boldsymbol{w} = (\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \cdots, \boldsymbol{w}^{(M)})$である．また，$m$ 番目の文書の n 番目の単語を $w_n^{(m)}$ で表す．m 番目の文書の単語数を N_m とすると，$\boldsymbol{w} = (w_1^{(m)}, w_2^{(m)}, \cdots, w_{Nm}^{(m)})$である．

潜在変数として，単語 $w_n^{(m)}$のトピック $z_n^{(m)}$ の他に，$w_n^{(m)}$ がフレーズ末尾か否かを表す変数 $x_n^{(m)}$ を導入する．$x_n^{(m)} = 1$ の場合，$w_n^{(m)}$ がフレーズ末尾であり，$x_n^{(m)} = 0$ の場合，$w_n^{(m)}$ はフレーズ末尾ではない（次の単語もフレーズとして続く）ことを意味している．$x_n^{(m)}$ の値は，$(w_n^{(m)}, z_n^{(m)})$ 毎のベルヌイ分布に従う．つまり，$w_n^{(m)}$ がフレーズ末尾か否かはその語とそのトピックに依存して確率的に定まるものとする．なお，単語とトピックの組 (w, k) 毎のベルヌイ分布のパラメタを$(\zeta_0(w, k), \zeta_1(w, k))$ とし（$\zeta_0(w, k) = 1-\zeta_1(w, k)$），その事前分布をパラメタ (δ_0, δ_1) のディレクレ分布（ベータ分布）とする．

$x_{n-1}^{(m)} = 1$ の場合，LDA と同様に，文書 m のトピックの混合比 $(\theta_1^{(m)}, \theta_2^{(m)}, \cdots, \theta_K^{(m)})$（$K$ はトピック数）に従って，トピック $z_n^{(m)}$ が生成され，$z_n^{(m)} = k$ とすると，トピック k が持つ単語の出現確率分布に従って，単語 $w_n^{(m)}$ が生成され，これがフレーズの先頭語となる．トピック k が持つ単語の出現確率分布はパラメタ $(\phi_1^{(k)}, \phi_2^{(k)}, \cdots, \phi_V^{(k)})$ の多項分布（V は語彙数）である．$(\theta_1^{(m)}, \theta_2^{(m)}, \cdots, \theta_K^{(m)})$ の事前分布をパラメタ $(\alpha, \alpha, \cdots, \alpha)$ のディレクレ分布，$(\phi_1^{(k)}, \phi_2^{(k)}, \cdots, \phi_V^{(k)})$ の事前分布をパラメタ $(\beta, \beta, \cdots, \beta)$ のディレクレ分布とする．

$x_{n-1}^{(m)} = 0$ の場合は，確率 1 で前単語のトピックが $z_n^{(m)}$ に引き継がれ，$z_{n-1}^{(m)} = z_n^{(m)} = k$ とすると，トピック k の単語バイグラムに従って単語 $w_n^{(m)}$ が生成される．トピック k で前単語が w のときの単語の分布 $P(\cdot \mid z_n^{(m)} = k, w_{n-1}^{(m)} = w)$ はパラメタ $(\eta_1^{(k,w)}, \eta_2^{(k,w)}, \cdots, \eta_V^{(k,w)})$ の多項分布であり，$(\eta_1^{(k,w)}, \eta_2^{(k,w)}, \cdots, \eta_V^{(k,w)})$ の事前分布をパラメタ $(\gamma, \gamma, \cdots, \gamma)$ のディレクレ分布とする．

潜在変数も単語と同様以下のようにまとめた表記を用いる．

$$\boldsymbol{z} = (\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \cdots, \boldsymbol{z}^{(M)})$$
$$\boldsymbol{z}^{(m)} = (z_1^{(m)}, z_2^{(m)}, \cdots, z_{N_m}^{(m)})$$
$$\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(M)})$$
$$\boldsymbol{x}^{(m)} = (x_1^{(m)}, x_2^{(m)}, \cdots, x_{N_m}^{(m)})$$

ただし，各文書の文書末はフレーズ末であるから，$x_{Nm}^{(m)}$ は 1 に固定する．

上記生成モデルに従った文書集合 $\boldsymbol{w}$ と各単語のトピック $\boldsymbol{z}$ およびフレーズ末尾か否かを表す $\boldsymbol{x}$ が生成される確率は，以下のように表すことができる．

$$P(\mathbf{w}, \mathbf{z}, \mathbf{x} \mid \theta, \phi, \eta, \varsigma)$$
$$= \prod_{m=1}^{M} \left\{ P(w_1^{(m)}, z_1^{(m)}, x_1^{(m)} \mid \theta^{(m)}, \phi, \varsigma) \times \prod_{n=2}^{N_m} P(w_n^{(m)}, z_n^{(m)}, x_n^{(m)} \mid w_{n-1}^{(m)}, z_{n-1}^{(m)}, x_{n-1}^{(m)}, \theta^{(m)}, \phi, \eta, \varsigma) \right\}$$

$$P(w, z, x \mid \theta^{(m)}, \phi, \varsigma) = \theta_z^{(m)} \phi_w^{(z)} \varsigma_x^{(z,w)}$$

$$P(w2, z2, x2 \mid w1, z1, x1 = 1, \theta^{(m)}, \phi, \eta, \varsigma) = \theta_{z2}^{(m)} \phi_{w2}^{(z2)} \varsigma_{x2}^{(z2,w2)}$$

$$P(w2, z2, x2 \mid w1, z1, x1 = 0, \theta^{(m)}, \phi, \eta, \varsigma) = \begin{cases} \eta_{w2}^{(z2,w1)} \varsigma_{x2}^{(z2,w2)} & ; z2 = z1 \\ 0 & ; z2 \neq z1 \end{cases}$$

本来は，文書末尾である確率 λ を導入し，次式の右辺に$(1-\lambda)^{(Nm-1)}\lambda$ を掛けるべきであるが，推定には関係しないため省略している．

各単語のトピック $\boldsymbol{z}$ およびフレーズ末尾か否かを表す $\boldsymbol{x}$ の推定では，

$$(z_1^{(1)}, x_1^{(1)}), (z_1^{(1)}, x_1^{(1)}), \cdots, (z_{N_1}^{(1)}, x_{N_1}^{(1)}), (z_1^{(2)}, x_1^{(2)}), \cdots, (z_{N_M}^{(M)}, x_{N_M}^{(M)})$$

の順に Gibbs Sampling により潜在変数の値をサンプリングし，これを繰り返す．$n < N_m$ の場合の $(z_n^{(m)}, x_n^{(m)})$ の Gibbs Sampling で用いられる条件付き確率を図 1 に示す．図中のカウンタの意味は以下の通りである．

$n_{dz}(m, k)$: $z_n^{(m)}$ と $x_n^{(m)}$ を除いた場合の，文書 m 中のフレーズ先頭語のトピックが k であるフレーズ数．

$n_{zw}(k, w)$: $z_n^{(m)}$ と $x_n^{(m)}$ を除いた場合の，フレーズ先頭語が w でそのトピックが k である個数．

$n_{wzw}(w1, k, w2)$: $z_n^{(m)}$ と $x_n^{(m)}$ を除いた場合の，トピック k のフレーズ内で $w1$ の次の単語が $w2$ である個数．

$n_{zwx}(k, w, 1)$: $z_n^{(m)}$ と $x_n^{(m)}$ を除いた場合の，トピック k のフレーズ末尾が語 w である個数．

$n_{zwx}(k, w, 0)$: $z_n^{(m)}$ と $x_n^{(m)}$ を除いた場合の，トピック k のフレーズで語 w がフレーズ末尾でない個数．

$n = N_m$ の場合も同様の考え方で条件付き確率を求めることができるが，紙面の都合上割愛する．ただし，$x_n^{(m)}$ は 1 に固定することになる．

次に，トピックを表す単語またはフレーズの表示について説明する．2 節末で述べたように，本研究では，各トピックについて，そのトピックで特徴的な単語またはフレーズのみを表示する．このために，tf・idf に似た考えでトピックにおける語（単語またはフレーズ）の重要度を与え，その上位のものを表示する．トピック k における語 w の重要度 $score(w; k)$ を以下のように定義する．

$$score(w;k) = tf(w;k) \times \log\frac{K}{df(w)}$$

ただし，$tf(w; k)$ は語 w にトピック k が割り当てられている頻度，$df(w)$ は語 w に割り当てられているトピックの異なり数である．

5. トピック分析例とその評価

2 節と同様，「NTCIR-1: 情報検索/用語抽出研究用テストコレクション」より取り出した用語「機械翻訳」を含む 636 件のアブストラクトをクエリ「機械翻訳」で論文を検索した結果のアブストラクト集合と見なす．この文書集合に対して，提案手法に基づくトピック分析を行った．また，比較のために，前処理として，各対象文書を単語とフレーズの列に変換し，これを LDA によりトピック分析した．単語またはフレーズの列に変換する際に複合的表現の取り出しには，2 通りの手法を試みた．

前処理に用いた手法の一つは単語の遷移確率を利用したものである．この手法では，全データを用いて単語 bigram の遷移確率 $P(w_2|w_1)$ を計算し，$P(w_2|w_1)$ が設定した閾値を超えるものであれば，w_1 と w_2 の間にはフレーズの境界は無く，w_2 はフレーズ内で w_1 に続いているとする．今回は，抽出されるフレーズを目視で確認し，この閾値を 0.2 に設定した．

以下で用いるカウンタの定義は次の通り．

$$n_{dz}(m,k) = \left|\{i \mid i \neq n \,\&\, (x_{i-1}^{(m)} = 1 \text{ or } i = 1) \,\&\, z_i^{(m)} = k\}\right|$$

$$n_{zw}(k,w) = \left|\{(i,j) \mid i \neq n \,\&\, j \neq m \,\&\, (x_{i-1}^{(j)} = 1 \text{ or } i = 1) \,\&\, z_i^{(j)} = k \,\&\, w_i^{(j)} = w\}\right|$$

$$n_{wzw}(w1,k,w2) = \left|\{(i,j) \mid i \neq n \,\&\, i > 1 \,\&\, j \neq m \,\&\, x_{i-1}^{(j)} = 0 \,\&\, z_i^{(j)} = k \,\&\, w_{i-1}^{(j)} = w1 \,\&\, w_i^{(j)} = w2\}\right|$$

$$n_{zwx}(k,w,x) = \left|\{(i,j) \mid i \neq n \,\&\, j \neq m \,\&\, z_i^{(j)} = k \,\&\, w_i^{(j)} = w \,\&\, x_i^{(j)} = x\}\right|$$

$$sn_{dz}(m) = \sum_k n_{dz}(m,k)\ ,\ sn_{zw}(k) = \sum_w n_{zw}(k,w)\ ,\ sn_{wzw}(w1,k) = \sum_{w2} n_{wzw}(w1,k,w2)\ ,\ sn_{zwx}(k,w) = n_{zwx}(k,w,0) + n_{zwx}(k,w,1)$$

$(z_n^{(m)}, x_n^{(m)})$ の Gibbs Sampling で用いられる条件付き確率（ただし，$n < N_m$ の場合，C は正規化のための定数）：

(Case1) $n > 1$ & $x_{n-1}{}^{(m)} = 0$ の場合

$$P(z_n^{(m)} = k, x_n^{(m)} = 0 \mid \boldsymbol{w}, \boldsymbol{z}/z_n^{(m)}, \boldsymbol{x}/x_n^{(m)}) = \begin{cases} C \cdot \dfrac{n_{wzw}(w_{n-1}^{(m)},k,w_n^{(m)}) + \gamma}{sn_{wzw}(w_{n-1}^{(m)},k) + V\gamma} \cdot \dfrac{n_{zwx}(k,w_n^{(m)},0) + \delta_0}{sn_{zwx}(k,w_n^{(m)}) + \delta_0 + \delta_1} \cdot \dfrac{n_{wzw}(w_n^{(m)},k,w_{n+1}^{(m)}) + \gamma}{sn_{wzw}(w_n^{(m)},k) + V\gamma} & ;\ z_{n-1}^{(m)} = z_{n+1}^{(m)} = k \\ 0 & ;\ otherwise \end{cases}$$

$$P(z_n^{(m)} = k, x_n^{(m)} = 1 \mid \boldsymbol{w}, \boldsymbol{z}/z_n^{(m)}, \boldsymbol{x}/x_n^{(m)}) = \begin{cases} C \cdot \dfrac{n_{wzw}(w_{n-1}^{(m)},k,w_n^{(m)}) + \gamma}{sn_{wzw}(w_{n-1}^{(m)},k) + V\gamma} \cdot \dfrac{n_{zwx}(k,w_n^{(m)},1) + \delta_1}{sn_{zwx}(k,w_n^{(m)}) + \delta_0 + \delta_1} \cdot \dfrac{n_{dz}(m,z_{n+1}^{(m)}) + \alpha}{sn_{dz}(m) + K\alpha} \cdot \dfrac{n_{zw}(z_{n+1}^{(m)},w_{n+1}^{(m)}) + \beta}{sn_{zw}(z_{n+1}^{(m)}) + V\beta} & ;\ z_{n-1}^{(m)} = k \\ 0 & ;\ otherwise \end{cases}$$

(Case2) $n = 1$ or $x_{n-1}{}^{(m)} = 1$ の場合

$$P(z_n^{(m)} = k, x_n^{(m)} = 0 \mid \boldsymbol{w}, \boldsymbol{z}/z_n^{(m)}, \boldsymbol{x}/x_n^{(m)}) = \begin{cases} C \cdot \dfrac{n_{dz}(m,k) + \alpha}{sn_{dz}(m) + K\alpha} \cdot \dfrac{n_{zw}(k,w_n^{(m)}) + \beta}{sn_{zw}(k) + V\beta} \cdot \dfrac{n_{zwx}(k,w_n^{(m)},0) + \delta_0}{sn_{zwx}(k,w_n^{(m)}) + \delta_0 + \delta_1} \cdot \dfrac{n_{wzw}(w_n^{(m)},k,w_{n+1}^{(m)}) + \gamma}{sn_{wzw}(w_n^{(m)},k) + V\gamma} & ;\ z_{n+1}^{(m)} = k \\ 0 & ;\ otherwise \end{cases}$$

$$P(z_n^{(m)} = k, x_n^{(m)} = 1 \mid \boldsymbol{w}, \boldsymbol{z}/z_n^{(m)}, \boldsymbol{x}/x_n^{(m)}) = C \cdot \frac{n_{dz}(m,k) + \alpha}{sn_{dz}(m) + K\alpha} \cdot \frac{n_{zw}(k,w_n^{(m)}) + \beta}{sn_{zw}(k) + V\beta} \cdot \frac{n_{zwx}(k,w_n^{(m)},1) + \delta_1}{sn_{zwx}(k,w_n^{(m)}) + \delta_0 + \delta_1} \cdot \frac{n_{dz}(m,z_{n+1}^{(m)}) + \alpha}{sn_{dz}(m) + K\alpha} \cdot \frac{n_{zw}(z_{n+1}^{(m)},w_{n+1}^{(m)}) + \beta}{sn_{zw}(z_{n+1}^{(m)}) + V\beta}$$

図 1. Gibbs Sampling で用いられる条件付き確率

前処理に用いたもう一つの手法は, C-value を利用したものである．C-value は以下の式で定義される．

$$C\text{-}value(CN) = (length(CN) - 1) * (n(CN) - \frac{t(CN)}{c(CN)})$$

CN: 複合名詞
length(CN): *CN* の長さ
n(CN): *CN* の出現回数
t(CN): *CN* を含むより長い複合名詞出現回数
c(CN): *CN* を含むより長い複合名詞異なり数

今回は連接する全ての名詞をフレーズとして C-value を計算し，抽出されるフレーズを目視で確認し，C-value の値が 10 を超えるものを，フレーズとした．

2 単語からなるフレーズに関しては，前処理によりフレーズの抽出を行った場合のトピック分析の結果のトピックの割り当てと提案手法によるトピックの割り当ては大きな違いはないと考えられる．特に単語の遷移確率を利用する手法による前処理の場合は，提案手法と類似したフレーズの解析を行うため類似した結果になると予想される．

しかし，3 単語以上から構成されるフレーズのトピック分析の結果は前処理を施して LDA によりトピック分析した結果と提案手法によるトピック分析の結果は異なると予想される．3 単語以上から構成されるフレーズの文書集合における頻度がそれほど高くない場合，Gibbs Sampling によるトピックの割り当てが不安定となり，その結果，そのフレーズに関するトピックの割り当ては信頼性がない．一方，提案手法の場合は，フレーズとしての頻度が低くてもそれを構成する個々の単語の遷移が多い場合は，安定してトピックが割り当てられ，その結果，トピック分析の結果も信頼性が高くなると考えられる．

表 3　提案手法によるトピック分析

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
抽出	辞書	ユーザ	選択	表現	解析	翻訳	日本語	構造	・
情報	語	入力	関係	的	文	手法	英語	意味	,
単語	データ	システム	場合	処理	構文	型	1	記述	動詞
共起	パターン	変換	文章	自然-言語-処理	生成	学習	技術	言語	手話
自動	分析	実験-結果	知識-データ	文法	日本語	事例	2	概念	研究
等	作成	規則	分類	・	特徴	検索	結果	手法-提案	音声-認識
訳語-選択	収集	可能	決定	現在	中国語	翻訳-結果	3	・	対象
テキスト	システム	方法-提案	語義	内容	方式	翻訳-手法	観点	知識	全体
本-手法	定型	実現	考慮	慣用-表現	上	利用	対応	中間-言語	訳文-生成
利用	方法	言語-間	計算-機	例	形態素-解析	ルール	英文	表層	音声-言語
自動的	複合語	処理	意味-的	対話	文-生成	翻訳-システム	評価	目的	対応
コーパス	必要	開発	意味-解析	さ	実現	比較	会話	自然-言語	活用
問題	法	書き換え	英-日	応用	諸略	アルゴリズム	調査	木	使用
文書	化	出力	制約	並列	長文	場合	数	依存	統合
訳語	ニュース-文	問題-解決	複数	問題-点	助詞	従来	副詞	レベル	報告

表 4　遷移確率を利用した手法によるトピック分析

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
検索	手話	語	自然-言語	ユーザ	1	日本語	中間-言語	ルール	こと
高速化	発話	知識-データ	知識	英文	2	動詞	概念	事例	機械-翻訳
ユーザ-辞書	音声-言語	抽出	理解	機能	0	意味	言語	シソーラス	システム
EBMT	音声-認識	単語	技術	支援	3	中国語	開発	格-パターン	ため
時間	話し言葉	多義性	名詞	作成	9	韓国語	概念-辞書	木-構造	翻訳
普及	意図	自然-言語	これら	インターフェース	5	用言	語義	定型-表現	結果
MT	自動-翻訳	コーパス	考察	フェーズ	8	慣用-表現	独立	アルゴリズム	文
近年	会話	複合語	代名詞	英文作成	4	文	中心	プログラム	提案
言語-処理	認識	解消	機能	OCR	評価	助詞	開発され	木	利用
実施	音声	方法	処理	流れ	パターン	語	説明	学習	必要
ネットワーク	TRANS	中	意味-言語	複数	7	格-助詞	ユーザー	概要	よう
文法	翻訳-実験	文字	意味	支援-システム	向上	英語	多言語間-機械	表層-表現	処理
必須格	技術	検討	概念	ツール	0%	構造	機械的	同様	本稿
類型	出力	専門-用語	課題	書き換え	名詞	助動詞	方式	日本語-動詞	手法
項目	対話	獲得	ゼロ	文節	動詞	副詞	中国	検索	解析

表 5　C-value を利用した手法によるトピック分析

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
言語	1	音声	翻訳	中国語	翻訳	翻訳	辞書	評価	格
意味	パターン	認識	システム	並列	手法	文	方法	技術	選択
知識	3	手話	機械	名詞	こと	こと	抽出	プログラム	制約
概念	2	ため	翻訳-システム	モジュール	機械	日本語	こと	作成	記述
自然-言語	0	技術	こと	特徴	ルール	表現	訳語	能力	副詞
理解	主語	話し言葉	ユーザ	助動詞	ため	解析	単語	評価-基準	動詞
処理	9	会話	ため	検討	用例	機械	語	報告	用言
中間-言語	記事	文字	機能	モデル	利用	処理	ため	我々	フレーム
こと	英	音声-言語	文書	翻訳-システム	結果	構造	動詞	内容	要素
文脈	ニュース-文	自動	情報	語	実験	変換	意味	項目	意味
関係	もの	合成	結果	性質	対訳	文法	共起	試験	関係
発話	0%	報告	利用	数量-表現	処理	生成	利用	前置詞	例文
研究	7	実験-システム	開発	中日-機械	システム	英語	名詞	検証	アスペクト
よう	4	解析部	作成	詞	問題	ため	情報	部分	困難
生成	6	入力	よう	ウイグル語	適用	言語	関係	調査	文型

表 6　各手法によるトピック分析の結果の評価

	遷移確率による手法	C-value による手法	提案手法
評価○のトピック数	4	3	5
評価△のトピック数	2	1	2
評価×のトピック数	4	6	3
評点	5.0	3.5	6.0

トピック数を 10 としてトピック分析を行った．提案手法によるトピック分析の結果を表 3 に示す．また，前処理として単語の遷移確率を利用してフレーズへの変換をした後 LDA によりトピック分析した結果を表 4 に，同じく前処理として C-value を利用した場合のトピック分析の結果を表 5 に示す．フレーズは単語同士を“ - ”（ハイフン）で繋いだものとして表す．

これらの結果を比較するため，トピック毎に提示された語及びフレーズを見てそのトピック内容を推定できるかどうかを三段階（○：1.0，△：0.5，×：0.0）で評価し，評点としてその合計値を算出した．3 つの手法に対する評価結果を表 6 に示す．

また，「情報-検索」や「日-英-翻訳」等，日本語の場合トピックを決定づける主辞はフレーズの後ろにあるパターンが多い．その性質を利用し，逆順からフレーズ生成を行うという手法も試みた．その結果及び評価を表 7, 8 に示す．

これらの結果から，オリジナルの LDA の結果よりも本手法の結果がトピックの内容理解に寄与していることが分かった．

6. 考察と今後の展望

今回の実験から，トピック内容の理解のためにはフレーズ生成機構が有効であるという足がかりを得ることができた．また，順方向での適用実験と逆方向での適用実験では逆方向の結果がより良いものであった．これは 5 章で述べた日本語の主辞が後ろにあるパターンが多くみられたためだと思われる．一方で，「翻訳-システム」や「音声-認識」等，トピックを決定付ける主辞がフレーズの先頭にある場合もあり，今後はどちらの場合でも対応できるような仕組みを考える必要がある．

本稿では，評価対象が一文書集合のみであり，また実験の評価者も少なかった．今後は評価対象と評価者を増やし，さらに客観的な評価を行う予定である．

謝辞　本研究を行うにあたり，大学共同利用機関法人情報・システム研究機構（国立情報学研究所）より，「NTCIR-1: 情報検索/用語抽出研究用テストコレクション」を実験データとして提供いただきました．ここに謝意を示します．

参考文献

[1] Blei, D.M., Ng, A.Y., & Jordan, M.I.(2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.
[2] T. L. Griffiths & M. Steyvers (2004). Finding scientific topics. Proc. of the National Academy of Sciences of the United States of America, vol. 101 Suppl 1, pp. 5228-5235.
[3] Katerina Frantziy, Sophia Ananiadouy(1996): Extracting nested collocations. Proc. of the 16th conference on Computational linguistics (COLING'96), Vol.1, pp41-44.
[4] J-stage https://www.jstage.jst.go.jp/browse/-char/ja.
[5] 牧田健作，鈴木浩子，小池大地，宇津呂武仁，河田容英(2012): Wikipedia を知識源とする分野トピックモデルの推定と分析．研究報告データベースシステム, 2012-DBS-155, 11 号, pp1-11.
[6] Nicola Barbieri, Giuseppe Manco, Ettore Ritacco, Marco Carnuccio & Antonio Bevacquae(2013). Probabilistic topic models for sequence data. Machine Learning, Vol. 93, Issue 1, pp 5-29.

表 7　提案手法（逆方向）によるトピック分析

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
自然-言語	機械-翻訳	概念	構文-解析	英-日	辞書	専門-用語	中国語	ルール	意味
手話	処理	プログラム	形態素-解析	日-英	データ	体系化	中日	用例	韓国語
発話	翻訳	中間-言語	意味-解析	記事	対訳	誤り	日中	手法	名詞
理解	こと	設計	意味-属性	ユーザー	抽出	OCR	両言語	事例	用言
文脈	システム	素性-構造	失敗	英語-ニュース	方法	照合-検索法	分解	検索	訳語
音声	文	依存-構造	長文	我々	知識	限定され	家族-モデル	文字	動詞
音声-認識	ため	自然-言語	結合	使用され	単語	推敲	父親-モジュール	シソーラス	格-助詞
音声-言語	提案	一部	前後	システム	語	冠詞	二-文字	帰納的-学習	副詞
モデル	情報	記述	日-英	韓-日	対	MBT-3	漢字-列	学習	述語
目的-言語	本稿	検証	優先-順位	英	3	前編集-自動化	太郎-モジュール	適用	日本語
話し言葉	よう	管理	長文-分割	語彙	機械-翻訳	正規化	次郎-モジュール	遺伝的	日-韓
状況	生成	PIVOT	単語間	離合-詞	訳語	誤り-検出	花子-モジュール	高速化	選択
意味-言語	変換	統語-構造	諺	日	共起	ワードプロセッサ	常用-文型	EBMT	助動詞
解析部	利用	原-言語	係り-解析	抽出法	作成	チェック	連体節-主節	類似度	マーカー
暗喩	日本語	過程	論理的-並列	時制	2	E-	連体-修飾	有効性-確認	助詞

表 8　提案手法（逆方向）の評価

	提案手法（逆方向）
評価○のトピック数	6
評価△のトピック数	2
評価×のトピック数	2
評点	7.0