文書内の事象を対象にした潜在的トピック抽出手法の提案とその応用

北島 理沙† 小林 一郎‡
†お茶の水女子大学理学部情報科学科
‡お茶の水女子大学大学院人間文化創成科学研究科理学専攻
{g0720520, koba}@is.ocha.ac.jp

1 はじめに

近年，文書上の潜在的トピックを扱う機会が増え，LSI，pLSI，LDAなどの潜在的意味解析手法が利用されるようになってきた．しかしこれらの手法において，トピックが割り当てられるのは単語であり，単語間の依存関係は考慮されていない．そこで本研究では，文書上の各事象をイベントとして定義し[1]，文書をイベントの集合として扱うモデルを提案する．潜在的意味解析手法としては，潜在的ディリクレ配分法（LDA）を用い，トピックの割り当て対象を単語からイベントに変更する．そして，文書検索課題を通じて文書の潜在的トピックを正しく推定出来ているかを確認し，従来の単語にトピックを割り当てる手法と比較をすることで，提案手法の特性，性能を調べる．また，その応用として，近年盛んになっているクエリに特化した要約[1]を対象とし，提案手法を用いた要約文生成を示す．

2 関連研究

LDAにおいては，トピックの割り当て対象を単語列に変更したことによって，より柔軟なトピック割り当てが出来ることが報告されている[2]．また，単語の依存関係を考慮した素性を扱うことで，文書分類の精度が向上することが報告されている[3]．

単語に対してトピックを割り当てる場合，単語の出現頻度が等しい2つの文書は，その語の依存関係にかかわらず，同じトピック分布をもつと推定されてしまう．しかし，単語の出現頻度よりもむしろ語と語の関係性が文書を表わす特徴量として重要となる場合がある．例えば，評価分類をする場合では，何に対してどのような意見を持っているか，という情報が重要になると考えられる．以上のような理由に基づき，本研究ではイベントを単位としたトピック割り当てを提案する．

また，テキスト要約に関する研究としては，従来の基本的な重要文抽出法以外に潜在的意味解析手法を用いた手法が提案されている[4][5]．これらにおいては，対象が文書である場合と同様にして文のトピック分布が推定され，それに基づいた要約文が生成される．本研究においても，文の潜在的トピック抽出に基づく要約文生成に提案手法が有用であることを示す．

3 イベントに基づいたトピック推定

文書検索において，各文書は文書を構成する単語とその重要度の積からなる文書ベクトルとして表現され，その重要度は索引となる単語の出現頻度を用いることが多い．しかし本研究では，イベントという単位で文書を扱うとするため，各文書に対してイベントを抽出し，文書群全体について索引となるイベントを決め，そのイベントの出現頻度を要素としたイベント－文書行列を作成する．そして，それに基づいてトピック推定を行う．

3.1 イベントの定義

イベントとは，文書上に存在している事象のことを指し，何が起こったか，誰がどのように感じたか，などの出来事を表わすような単語の組として表現する．その抽出方法について述べる．

まず，文書に対して構文解析器CaboCha[2]を用いて文節の係り受け関係を取り出す．そして，係り受け関係にある2つの文節から単語を抽出し，(主語，述語)，(述語1，述語2)の条件を満たす組をイベントと定義する．主語には名詞，未知語が，述語には動詞，形容詞，形容動詞がそれぞれ該当する．(述語1，述語2)をイベントとして選んだ理由は，予備実験にて実際に

[1]イベントの定義については，3章で詳述する．

[2]http://chasen.org/ taku/software/cabocha/

抽出されたイベントと文書を見比べることによりその必要性を確認したこと，および，主語が省略されている文に対しては前者のタイプのイベントが抽出できないことによる．

3.2 イベント－文書行列の作成

通常，単語－文書行列を作成する際に，どのような文書においても一般的に頻出する単語と，文書群において極端に出現頻度の少ない語は除去されることが多い．提案手法では，予備実験において前者のような除去すべき頻出イベントは見受けられなかった．これは，イベントという単語の組にすることで必要性の低い語にも意味が付与され，どれも文書を特徴づける素性として扱う必要が出てくるためであると考えられる．一方，後者のような出現頻度の少ないイベントは非常に多く見受けられた．このことは，イベントの性質から明らかであり，素性の持つ意味が単語の場合と異なるため，同様の処理では対応できない場合が存在する．具体的には，文書群において出現頻度が 1 であるイベントを全て除去してしまうと，文書内容の再現性の低い文書ベクトルが生成されてしまうことがある．このことを踏まえ，それを除去してしまうと文書ベクトルの要素が消えてしまうようなイベントは，たとえ出現頻度が 1 であっても残し，文書としての再現性を保つことにする．

3.3 トピック分布の推定

イベント－文書行列の作成後，潜在的ディリクレ配分法 [6] によってトピック推定を行う．潜在的ディリクレ配分法とは，一つの文書に対して複数のトピックが存在すると想定した確率的トピックモデルであり，それぞれのトピックがある確率を持って文書上に生起するという考えの下，そのトピックの確率分布を導き出す手法である．各トピックは，語彙の多項分布として表現される．

本研究では，トピックの割り当て対象はイベントとなるため，各トピックはイベントの多項分布として表現される．トピック推定に用いる手法としては，変分ベイズ法 [6] などが提案されているが，本研究ではギブスサンプリングによる推定 [7] を行うこととする．また，クエリのトピック分布については，クエリに含まれる各イベントの持つトピック分布の総和とする．

4 文書検索による性能評価実験

共通の文書検索課題を通じて，従来手法と提案手法の性能を比較および評価する．具体的には，クエリの持つトピック分布と類似するトピック分布を持った文書を検索結果とし，検索結果の精度を調べることで，推定されたトピック分布が各文書の意味を捉えられているかを確かめる．以後，従来手法を “wordLDA”，提案手法を “eventLDA” と呼ぶ．

4.1 トピック分布類似度判定指標

トピック分布の類似度判定指標としては，Kullback-Leibler 距離，Symmetric Kullback-Leibler 距離，Jensen-Shannon 距離，cosine 類似度を用いて比較を行う．wordLDA においては，Jensen-Shannon 距離を用いたときが最も精度が高いと報告されており [5]，提案手法でも同様にして比較を行うことにする．Kullback-Leibler 距離を D_{KL} で表わすとき，Symmetric Kullback-Leibler 距離，Jensen-Shannon 距離は，それぞれ式 (1)，式 (2) で定義される．

$$D_{symKL}(S,Q) = D_{KL}(S \parallel Q) + D_{KL}(Q \parallel S) \quad (1)$$

$$D_{JS}(S,Q) = \tfrac{1}{2}D_{KL}(S \parallel M) + \tfrac{1}{2}D_{KL}(Q \parallel M) \quad (2)$$
$$M = \tfrac{1}{2}(S+Q)$$

4.2 実験仕様

対象データには，楽天トラベル[3]のホテル・施設に関する評価・レビューを用いた．レビューには，「部屋」や「立地」などの各対象につき 1〜5 の 5 段階評価があり対象と評価の関係性が保持されているため，提案手法の性能評価に適していると考える．クエリは「部屋が良かった」とし，対象文書群は「部屋」の評価が 1 のレビューから無作為に選んだ 1000 件，5 のレビューから無作為に選んだ 1000 件の合計 2000 件とする．正解文書は，評価が 5 のレビュー 1000 件である．評価指標には，11 点平均適合率を使用する．

本実験では，適切なトピック数と有効な類似度判定指標の調査の 2 つの観点から両手法の比較を行う．まず，類似度判定指標を Jensen-Shannon 距離に固定し，トピック数 k を $k = 5, 10, 20, 50, 100, 200$ と変化させる．次に，トピック数を先の実験によって得られた値に固定し，類似度判定指標を変化させる．ギブスサンプリングの反復回数は 200 回，各条件における試行回数は 20 回として，その平均をとる．wordLDA についても同様の実験を行い，その結果を提案手法と比較する．

4.3 実験結果

表 1 に，トピック数 k を変化させたときの 11 点平均適合率を示す．eventLDA では $k = 5$ のとき，wordLDA

[3] http://travel.rakuten.co.jp/

では $k = 50$ のときに精度が最も高くなっている．また，全体的にも eventLDA は wordLDA に勝る精度を保っていることが分かる．

表 2 に，類似度判定指標を変化させたときの 11 点平均適合率を示す．どの指標を用いた場合でも，eventLDA は wordLDA に勝る精度を保っていることが分かる．また，最も高い精度を示した指標については，wordLDA では Jensen-Shannon 距離，eventLDA では cosine 類似度となっている．逆に精度が低くなるのは両手法とも Kullback-Leibler 距離と共通であった．

表 1: トピック数による比較

トピック数	wordLDA	eventLDA
5	0.5152	0.6256
10	0.5473	0.5744
20	0.5649	0.5874
50	0.5767	0.5740
100	0.5474	0.5783
200	0.5392	0.5870

表 2: 類似度判定指標による比較

類似度判定指標	wordLDA	eventLDA
Kullback-Leibler 距離	0.5009	0.5056
Symmetric Kullback-Leibler 距離	0.5695	0.6762
Jensen-Shannon 距離	0.5753	0.6754
cosine 類似度	0.5684	0.6859

4.4 考察

実験結果より，提案手法は従来手法に比べて高い性能を示しており，文書の内容をより細かく捉えたトピック推定が行えていることが分かった．また，提案手法の特性として，少ないトピック数で分類が行えていることが分かった．その理由として，各素性の持つトピックがある程度狭い範囲に絞られ，結果として，誤差であるトピックが生成されないのではないか，と考える．

一方で，提案手法における最適な類似度判定指標は cosine 類似度となり，確率分布の類似度判定指標として用いられている指標の方が精度が低くなるという，予想に反した結果となった．このことから，トピック分布の確率分布としての性質についても調査が必要であると考える．

5 テキスト要約への応用

提案手法の応用例として，複数文書を対象としたテキスト要約を行う．クエリに特化した要約を行い，生成された要約文の精度を従来手法と比較することで提案手法の有効性を示す．

5.1 MMR-MD に基づく重要文抽出

クエリとの類似度のみを考慮すると冗長性のある要約文が生成される可能性があるため，それを防ぐための MMR-MD という指標が提案されている [8]．これは，既に抽出された文との類似度をペナルティとして与えることで，内容の重なる文の抽出を妨げる指標であり，式 (3) で定義される [9]．本研究では，潜在的トピックに基づいてクエリとの類似度が高い文を選びつつ，表層的には冗長性を削減することを目指し，クエリとの類似度判定 Sim_1 にはトピック分布の類似度を用い，既に抽出された文との類似度判定 Sim_2 には素性を単位とした cosine 類似度を用いる．

$$MMR\text{-}MD \equiv argmax_{C_i \in R \setminus S}[\lambda Sim_1(C_i, Q) - (1-\lambda) max_{C_j \in S} Sim_2(C_i, C_j)] \quad (3)$$

C_i ： 文書集合中の文
Q ： クエリ
R ： 文書集合からクエリ Q によって検索された文集合
S ： R の内，既に重要文として抽出されている文集合
λ ： 重み調整パラメータ

トピック分布の類似度判定指標としては，前章での実験で使用した 4 つの指標を用いて比較する．なお，$\lambda = 0.5$ とした．

5.2 実験仕様

本実験では，NTCIR4 TSC3[4] で用いられたテストセットを利用する．約 10 記事から成る文書セットが 30 トピック分用意され，総文数は 3587 文である．評価のために用意された質問集合を 1 つのクエリとし，クエリに特化した要約文生成の課題と見なす．評価方法としては，TSC3 において用いられた Precision と Coverage を使用し [10]，抽出する文数は，TSC3 で定められた文数とした．また，本手法の特性についても調べるために，トピック数，類似度による比較を行う．各条件につき試行回数は 20 回とし，30 文書セット中，無作為に選んだ 5 セットについて同様の実験を行い，平均をとる．提案手法の比較対象として，MMR-MD を評価指標として wordLDA を用いた場合の実験も行う．

5.3 実験結果

類似度判定指標による差は現れず，どの指標を用いた場合も同一の結果となった．表 3 に wordLDA と eventLDA による Precision と Coverage の比較を示す．最も精度の高いトピック数 k については，wordLDA では $k = 5$，eventLDA では $k = 10$ となっている．

[4] http://research.nii.ac.jp/ntcir/index-en.html

表 3: トピック数による比較

トピック数	wordLDA		eventLDA	
	Precision	Coverage	Precision	Coverage
5	0.314	0.249	0.404	0.323
10	0.264	0.211	0.418	0.340
20	0.261	0.183	0.413	0.325
50	0.253	0.171	0.392	0.319

さらに，潜在的意味解析を利用しないテキスト要約手法との比較を表 4 に示す. 文書を時系列順に並びかえ各文書の先頭から順に 1 文ずつ重要文として抽出する手法である Lead 手法，TF-IDF に基づいた重要文抽出手法を比較対象とし，それらの精度に関しては，先行研究 [10] で示されている実験結果の値を用いた.

表 4: 手法間の比較

手法	Precision	Coverage
Lead	0.426	0.212
TF-IDF	0.454	0.305
wordLDA (k=5)	0.314	0.249
eventLDA (k=10)	0.418	0.340

5.4 考察

どの条件においても eventLDA は wordLDA より高い精度を示し，提案手法は文に対するトピック割り当てにも有効であることが分かった．また，その精度が類似度判定指標によらない理由として，推定されたトピック分布が偏った分布となっており，指標による影響が現れなかったのではないかと考える．さらに，適切なトピック数は eventLDA の方が大きくなっており，新聞記事群を対象としたことから 1 つの単語に対するトピックがある程度決まっていたため，wordLDA では少ないトピック数で分類が行えたと考える．また，他の手法との比較において，提案手法はそれらと近い精度を示しており，表層的な情報を直接扱った場合と同じ程度の性能を持つことが分かった．特に，Coverage においては高い精度を示しており，潜在的トピックを扱ったことでより網羅的な要約文生成が行えたと考える.

6 おわりに

本研究では，係り受け関係に基づいた 2 つの単語の対をイベントと定義し，イベントにトピックを割り当てることで文書内の事象を捉えた潜在的トピック抽出手法を提案した．そして，4 章において提案手法の性能について調べ，その応用として，5 章では提案手法を用いたテキスト要約を示した.

対象が文書であっても文であっても，提案手法である eventLDA は，wordLDA よりも高い性能を持っていることを示すことができ，トピックをイベントという単位に割り当てた場合でも潜在的なトピックが推定できていることが分かった．イベントは，2 つの単語の関係性を保持することができるため，単語にトピックを割り当てる場合よりも様々な文書データに応用が可能である．本研究によって，素性をイベントのような情報量の大きいものにした場合でも潜在的なトピックを推定できることが分かり，単語以外の素性の有効性も示すことができた.

今後は，様々なタイプのデータ，クエリを用いて実験を行い，提案手法の特性についてさらに考察を行うつもりである．また，MMR-MD における重み調整パラメータ λ についても，様々な値で実験を行い，提案手法の特性に適した値について検証を行いたい．さらに，対象を文とした場合においては，抽出イベントの少なさの影響が大きくなることが考えられ，イベント抽出方法や除去すべきイベントの決め方などについてより深く考察を行っていくつもりである.

謝辞

本研究では，楽天技術研究所の許諾を頂き “楽天トラベル” のデータを利用させて頂きました．ここに深く感謝の意を表します.

参考文献

[1] 桜井 俊彦，内海 彰，情報検索のためのクエリに基づく文書自動要約，言語処理学会第 10 回年次大会発表論文集，pp, 265-268, 2004.
[2] 鈴木 康広，上村 卓史，喜田 拓也，有村 博紀，潜在的ディリクレ配分法の単語列への拡張，データ工学と情報マネジメントに関するフォーラム，2010.
[3] 松本 翔太郎，高村 大也，奥村 学，単語の系列及び依存木を用いた評価文書の自動分類，第 3 回情報科学技術フォーラム (FIT 2004) 講演論文集，2004.
[4] Q. Bing, L Ting, Z. Yu, and L. Sheng, Research on Multi-Document Summarization Based on Latent Semantic Indexing, Journal of Harbin Institute of Technology, 12(1): 91-94, 2005.
[5] L. Henning, Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis, International Conference RANLP 2009-Borovers, Bulgaria, pp. 144-149, 2009.
[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
[7] T. Grififths and M. Steyvers, Finding scientific topics, In Proc. of the National Academy of Sciences, Vol. 101, pp. 5228-5235, 2004.
[8] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, Multi-document summarization by sentence extraction, In Proc. of ANLP/NAACL Workshop on Automatic Summarization, pp. 40-48, 2000.
[9] 奥村 学，難波 英嗣，知の科学　テキスト自動要約，オーム社，2005.
[10] 平尾 努，奥村 学，福島 孝博，難波 英嗣，TSC3 コーパスの構築と評価，言語処理学会年次大会発表論文集，10th, A10B5-02, 2004.