文書上の事象を対象にした潜在的トピック抽出手法の提案

北島 理沙　(指導教員：小林 一郎)

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，これらにおいてトピックは単語に割り当てられ，単語間の依存関係については考慮されていない．これに対し，トピックの割り当て対象を単語列に変更することで柔軟なトピック割り当てができることなどが報告されている [1].

本研究では，文書上の各事象を単語の対で表現した“イベント” として定義し，LDA においてトピックの割り当て対象を単語からイベントに変更した手法を提案する．そして，文書検索課題と要約文生成課題を通じて従来手法と比較を行い，提案手法の性能について調べる．

2　イベントに基づいたトピック推定

2.1　イベント－文書行列

本研究における “イベント” とは，文書上に存在す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 2 つの単語の組として定義する．

イベントの抽出方法として，構文解析器 CaboCha[1] を用いて各文書から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，形容動詞が該当する．

従来手法では，単語－文書行列を作成する際に一般的な頻出語と出現頻度の極端に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頻出イベントは見受けられなかった．一方，後者のような出現頻度の極端に少ないイベントは非常に多く見受けられた．これらに対して従来手法と同様に全てを除去してしまうと，文書としての再現性が失われてしまうことがあると考えられる．従って，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ったイベント－文書行列を作成し，それに基づきトピック推定を行う．

2.2　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 [2] 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，各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，その確率分布を導き出す手法である．各トピックは，従来手法では語彙の多項分布として表現されるが，提案手法では，イベントの多項分布として表現される．本研究では，トピック推定手法としてギブスサンプリングを用いる．また，クエリのトピック分布は，クエリ上の各イベントの持つトピック分布の総和とする．

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

3　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，そ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．

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 類似度を用いる．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 [3]，提案手法でも同様にして各指標による比較を行う．

3.1　実験仕様

対象データには，楽天トラベル[2]のホテル・施設に関する評価・レビューを用いた.「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すると考える．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，トピック数と類似度判定指標の 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で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．

3.2　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 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 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 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，

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

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 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

3.3 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，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．今後，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．

4 テキスト要約による性能評価実験

対象を文とした場合の性能評価実験として，複数文書を対象とする，クエリに特化した要約文生成を行う．要約文生成において，クエリとの類似度のみを考慮すると冗長な要約文が生成される可能性があり，それを防ぐため MMR-MD という指標が提案されている [4]．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る．本研究では，クエリとの類似度判定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には素性を単位とした cosine 類似度を用いる．

4.1 実験仕様

本実験では，NTCIR4 TSC3[3] で用いられたテストセットを利用する．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課題と見なす．3 章の実験と同様に，トピック数，類似度判定指標による比較をし，TSC3 で用いられた Precision と Coverage により評価を行う．各条件につき試行回数は 20 回とし，平均をとる．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．

4.2 実験結果

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 の比較を示す．wordLDA では $k = 5$，eventLDA では $k = 10$ のとき，精度が最も高くなっている．

さらに，潜在的な意味を考慮しない要約手法との比較を表 4 に示す．各文書の先頭から順に重要文として抽出する Lead 手法，TF-IDF に基づいた重要文抽出手法を比較対象とした．

[3]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

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

4.3 考察

どの条件においても eventLDA は wordLDA より高い精度を示し，提案手法は文に対して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．

5 おわりに

本研究では，係り受け関係に基づいた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．対象が文書であっても文であっても提案手法は高い性能を持つ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トピックが推定できることが分かった．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．

参考文献

[1] 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，第 2 回データ工学と情報マネジメントに関するフォーラム，2010.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan : Latent Dirichlet Allocation, Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.

[3] L. Henning : Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis, Proc. of the International Conference RANLP-2009, pp.144–149, 2009.

[4] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz : Multi-document summarization by sentence extraction, Proc. of ANLP/NAACL Workshop on Automatic Summarization, vol.4, pp.40–48, 2000.