4R-3

文書内のイベントを対象にした潜在的ディリクレ配分法によるクエリに特化した要約

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情報科学コース

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，その依存関係については考慮されていない．これに対し著者らは，文書中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（本稿では，“wordLDA”と呼ぶ）からイベントに変更した手法（“eventLDA”と呼ぶ）を提案した [1]．本研究では，近年盛んになっているクエリに特化した要約を対象とし，[1] にて提案した手法を応用した要約文生成を示す．

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[2] 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，形容動詞がそれぞれ該当する．

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 [3] 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．

3 テキスト要約への応用

複数文書を対象とした，提案手法 [1] によるテキスト要約について述べる．

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り，式 (1) で定義される [5]．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．

Query-biased Summarization using Latent Dirichlet Allocation based on Events in a Document

†Risa KITAJIMA(g0720520@is.ocha.ac.jp),
‡Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
†Dept. of Information Sciences, Faculty of Science, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610
‡Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University,2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

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

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，Kullback-Leibler 距離，Symmetric Kullback-Leibler 距離，Jensen-Shannon 距離，cosine 類似度を用いる．なお，λ=0.5 とした．

4 実験

4.1 実験仕様

本実験では，NTCIR4 TSC3* 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 [6]，抽出する文数は，TSC3 で定められた文数とした．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．

4.2 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 1 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．

表 1: 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 2 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 [6] で示されている実験結果の値を用いた．

*http://research.nii.ac.jp/ntcir/index-en.html

表 2: 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，文書内のイベントを対象に潜在的なトピックを割り当てる手法 [1] を用い，トピック抽出対象を文とした要約文生成を行った．これにより，提案手法による潜在的な意味に基づいた要約文生成の有効性を示した．今後は，様々なタイプのデータを用い，MMR-MD における重みパラメータの調整を行うなどして，提案手法の特性を活かした要約手法について考察を行っていくつもりである．

参考文献

[1] 北島 理沙，小林 一郎，文書内のイベントを対象にした潜在的トピック抽出手法の提案，情報処理学会第 73 回全国大会，2S-2，2011.
[2] http://chasen.org/ taku/software/cabocha/
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
[4] 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.
[5] 奥村 学，難波 英嗣，知の科学 テキスト自動要約，オーム社，2005.
[6] 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，言語処理学会第 10 回年次大会発表論文集，A10B5-02，2004.