LDAによる国会会議録を対象にしたセグメンテーションの一手法

伊藤誉将† 関洋平† 青野雅樹†
豊橋技術科学大学†

1 はじめに

インターネットの発展に伴い，膨大なデータの自由な閲覧や取得が可能となった．その中でも大半を占めるテキストデータは，量の多さだけでなく，形式や長さがまちまちである．長いテキストでは，これを短いまとまりに区切り，読みやすくすることが重要である．そこで，トピックの変わり目ごとに分割(セグメンテーション)し短いまとまりにする．このようにすることで情報を得ようとする側の負担を軽減できると期待できる．

本研究では，長いテキストの代表例として国会会議録[4]に着目し，トピック発見に有効とされるLDA(Latent Dirichlet Allocation)[3]を用いた会議録のセグメンテーションを行う手法を検討する．

2 関連研究

確率的トピックモデルを用いた研究は近年数多くなされている．中村ら[1]はテキストに対し文脈長を最適化しつつ LDA のオンライン適応を行っている．津田ら[2]は隣接ブロック間の距離により文脈長の制御を行いトピック変化点の抽出を行っている．これらの研究では，異なるトピックのニュースを複数接続したテキストを対象としている．

本研究では一つのテキストで複数のトピックを含む国会会議録を用い LDA を適用し，セグメンテーションを行った．

3 LDA

LDA は，潜在的なトピック生成確率がディリクレ分布に従うと仮定したモデルである．本研究におけるテキスト d の生成確率は以下の(1)式で定義される．

$$P(d\mid\alpha,\beta)=\int p(\theta\mid\alpha)\left(\prod_{n=1}^{|d|}\sum_{z_k}p(w_n\mid z_k,\beta)p(z_k\mid\theta)\right)d\theta \quad (1)$$

θ ：トピック出現確率

z_k ：潜在的トピック

w_n ：テキスト d 内に含まれる単語 $\left(1\le n\le|d|\right)$

また，α と β は LDA のモデルパラメーターである．α はディリクレ分布のパラメーターであり，β はあるトピック z において語 w が出現する確率を示している．

A Method of Segmentation intended for the Diet Minutes by LDA(Latent Dirichlet Allocation)
†Takamasa Ito, Yohei Seki, Masaki Aono
†Toyohashi University of Technology

4 国会会議録のセグメンテーション

本研究では LDA を用いて教師なし学習を行う．学習データ及びテストデータは国会会議録であるため，それらの作成を行なう．データは「第 171 回国会衆議院予算委員会」全 28 会議録を対象とした．

4．1 国会会議録の発言者情報

国会会議録では参加者一覧や各発言の文頭から“○前原委員”等といった人物（発言者）情報を得ることができる．この人物情報を利用し発言単位にまとめる．しかし，他のテキストにも適用を考慮しているため実際の発言内に出現する人物名のみ利用する．また，ある発言者から次の発言者の直前までを発言とする．

4．2 発言の接続処理

国会会議録では質問とそれに対する答弁が順に出現する．そのためこれらを組にすることも考えたが，質問者と答弁者の他に司会者の発言があり上手くまとまらない事がある．また，司会者発言に多く見られるのだが発言自体が短い，特徴語抽出を行った際に一つも抽出されないといった問題が生じた．そこで，以下のルールに基づき発言を人手により接続した．

1．発言の文字数が N 文字以下であれば以降の処理を行う．
2．司会者発言であれば一つ前の発言に接続する．
3．質問であれば答弁に，答弁であれば質問に接続する．
4．ただし，連続して N 文字以下の発言が続く場合は前の発言に接続し続ける．
5．1～4の処理を N 文字以下の発言が無くなるまで繰り返す．

4．3 特徴語抽出

特徴語抽出を行い，これらを要素とする学習及びテストデータ用の文書ベクトルを作成する．

特徴語抽出には形態素解析を行い，名詞と未知語を特徴語候補とする．これら全てを特徴語とはせずに名詞，未知語が連続する間は接続し続ける複合語作成の処理及び，出現頻度の制限加える．また，代名詞などの単独では意味をなさない語はストップワードとして利用しない．この処理を加えることで複合語かつ特徴的な語を得ることができると考えたためである．

学習データ作成時は上記の処理のみだが，テストデータ作成時さらに処理を加える．例えば，一発言内に“労働条件”と“条件”が同時に現れた場合，この“条件”は“労働条件”を指している

と仮定し同じものする．これにより発言内の語の繋がりを詳細に見ることができると考えたためである．しかし，これらの語を平等に扱うのは語から得られる情報の差が大きいと考えられるため以下の(2)式を用いて計算する．

$$wtf(w_n) = \mathrm{round}\left(\sum_{j=1}^{J} \frac{size(word_j)}{size(w_n)} \times tf(word_j)\right) \quad (2)$$

ここで，$word_j$ は w_n の部分文字列であることを指している．

4．4　トピックの変わり目の判定方法

LDA による学習を T 試行行い，得られた全ての α と β 及びテスト用の文書ベクトルを用いて変分ベイズ法(3)式に適用し収束するまで繰り返し行う．

$$\phi(n,i) = \beta(i,w_n)\exp\left(\psi(\gamma_i) - \psi\left(\sum_{h=1}^{k}\gamma_h\right)\right) \quad (3)$$

$$\gamma_i = \alpha(i) + \sum_{n=1}^{N} wtf(w_n)\phi(n,i)$$

ここで得られた γ_i が一つの発言が持つトピックの割合を示している．この結果を用いて発言間の類似度計算を行う．

T 試行分の類似度計算の結果より，類似度が低いものから正解数だけ取り出しトピックの変わり目の候補とする．次に T 試行中何回候補として取り出されたかを確認し，$T/2$ 回以上取り出されたときのみトピックの変わり目として抽出する．この作業により抽出されるトピックの変わり目の総数(抽出数)は正解数だけ取り出せると限らない．

5　実験

5．1　実験概要

国会会議録に対する LDA を用いたセグメンテーションの有効性を確認するため評価実験を行った．また，セグメンテーションの対象として国会会議録「第 171 回国会衆議員予算委員会 20 号」をテストデータ，それ以外を学習データとして使用した．

5．2　実験方法

LDA で学習するデータは $N=100$，$T=8$ とした時，4節で述べた方法で予め文書ベクトル化したものを利用する．また本実験では LDA のトピック数および特徴語の出現頻度に変化をつけ実験を行った．

5．3　正解セットの作成

会議録の内容が変わっているかどうかを判定するには実際に発言の前後の内容が別の内容かどうかを判断する必要がある．そのため予め人が読んでトピックの変わり目を発言間で抽出し正解セットとした．加えて，国会会議録の性質上，一発言内でトピックが変わることがある．この場合，本研究では一発言の前後の発言間両方でトピックの変わり目と判定した．また，“第 171 回国会衆議員予算委員会 20 号”のトピックの変わり目は 54 個で，これが正解総数である．

5．4　評価

表1　各条件における正解数

特徴語頻度	特徴語数	トピック数	抽出数	正解数
1～5	23700	30	60	18
		50	58	16
2～5	7856	30	55	9
		50	52	9
1～10	25443	30	55	13
		50	57	15

特徴語の出現頻度とトピック数毎の抽出数，正解数を表 1 に示す．特徴語頻度は特徴語の出現頻度の制限(最小値～最大値)でありこれにより得られた特徴語の総数を特徴語数，また抽出数のうちいくつ正解であったかを正解数で示している．表1より総じて低い結果であるが，特徴語頻度数を小さい値に制限することで国会会議録のセグメンテーションに有効であることがわかった．

5．5　考察

ベクトル作成時の発言接続と特徴語抽出及び対象としたデータについて考察を行う．

特徴語頻度では，国会会議録内によく出現する頻度が高いものより出現頻度が低い語を利用することが有効であることがわかった．このような語は，LDA による語のトピック分類時にある一つのトピックにしか存在しない語となり他のトピックへの影響を及ぼしていないためと考えられる．

発言の接続処理によりトピックの変わり目が内部に含まれることはなかったが，他のトピックに分類されたであろう語の存在は確認された．

本研究ではニュースに対して行われた実験を応用して行った．対象とした国会会議録では一般的なニュースやブログと違い，言葉のやり取りが行われている特殊なテキストである．そのため，特徴語抽出や発言接続といった処理のみでは低い結果になったと考えられる．

6　おわりに

国会会議録のセグメンテーションをトピック発見に用いられる LDA を用いて行った．結果から成果は低かったが，出現頻度が少ない語は LDA による国会会議録を対象としたセグメンテーションに有効であるとがわかった．今後は，発言間の関係を利用した特徴語抽出を行うことがあげられる．

参考文献

[1] 中村明，速水悟，津田裕亮，松本忠博，池田尚志．“トピック変化点に基づく LDA のオンライン適応における複数モデル統合の効果” 言語処理学会第 15 回年次大会，2009

[2] 津田裕亮，中村明，速水悟，松本忠博，池田尚志．“LDA トピックモデルに基づく話題変化点検出” 言語処理学会第 15 回年次大会，2009

[3] D. Blei, A. Y. Ng and M. Jordan, “Latent dirichlet allocation” Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[4] 国会会議録検索システム，国立国会図書館，Available from http://kokkai.ndl.go.jp/