文書から取得した制約に基づく潜在トピック抽出

立川華代　(指導教員：小林 一郎)

1　はじめに

近年，膨大な量の文書を処理する機会が増えてきている．大量の文書について，表層的な情報に基づく処理ではなく，Latent Dirichlet Allocation(LDA) を用いて，文書中の潜在的なトピックを抽出をする研究が多く行われてきている．LDA を使ってトピックを抽出する際に，人の判断において同一トピックに入ると想定される語が同一トピックのものとはならないことがある．それに対して，文章に含まれる単語から同一トピックに入るべきだと考えられる単語群を選択し，それらが同一トピックに入るように事前知識として制約を与える手法が提案されている [1, 2]．しかし，それらの制約はユーザの主観によって与えられるケースが多く，対象となる文書から制約となる知識を自動で取得しているわけではない．そこで本研究では，与えられた文書から制約となる単語群を自動的に抽出し，それらを事前知識として与えることで制約を踏まえた潜在トピック抽出を行い，その結果を考察する．

2　関連研究

Andrzejewski ら [1] は，Must-Links と Cannot-Links の二つの制約を設定し，ディリクレ森分布を用いて，トピック分類される単語に制約を付与した．また，Kobayashi ら [3] は，Andrzejewski ら [1] が採用した Must-Links と Cannot-Links において論理的演算による制約知識の結合を利用できるようにし，論理的な制約に基づき新たに制約を追加することも可能にするトピック抽出手法を提案している．

一般的な制約付きクラスタリングでは，制約は人手で与えられることが前提となっていることから，鍛冶ら [4] は，語彙統語パターンを用いて約 10 億文のコーパスから類義語を自動獲得し，それに基づいた制約を構築することにより制約付きの単語クラスタリング手法を提案している．本研究では，そのような大きな制約知識を前提とせず，潜在トピックを分類する対象文書から制約知識を抽出することにより，トピック分類の精度が向上できるかを検討する．

3　事前知識に基づくトピック抽出

3.1　Dirichlet Forest LDA

LDA を利用し，制約を組み込んで潜在トピックの分類を行うために，ディリクレ分布にディリクレ森分布（Dirichlet Forest Prior，以下 DF）を用いる．DF とはディリクレ分布を階層化したものであり，通常の LDA と同様にディリクレ分布のハイパーパラメータとして α と β をそれぞれトピック分布と単語出現分布に用いるが，それに加え単語出現分布において，与えた制約の強さを反映する η を用いる．DF を用いた LDA（以下，LDA-DF）では d_i，z_i を i 番目の単語 w_i が含まれる文書および割当てられるトピックとし，上述したパラメータを用いて以下の式で表現される．

$$\theta_{d_i} \sim \text{Dirichlet}(\alpha) \quad (1)$$

$$z_i|\theta_{d_i} \sim \text{Multinomial}(\theta_{d_i}) \quad (2)$$

$$q \sim \text{DirichletForest}(\beta, \eta) \quad (3)$$

$$\phi_{z_i} \sim \text{DirichletTree(q)} \quad (4)$$

$$w_i|z_i, \phi_{z_i} \sim \text{Multinomial}(\phi_{z_i}) \quad (5)$$

3.2　制約知識の構築

制約知識を構築するのに，トピックを代表すると見なせる単語を選択する必要がある．本研究においては，あるトピックを代表する単語（以下,「重要単語」と呼ぶ）とは，対象とする文書群に万遍無く高頻度に現れるもの，または，他の単語と多くの共起関係にあるものと仮定するため,「頻度情報」と「共起情報」をもとに重要単語を抽出する．以下のステップに従い，制約知識の構築を行う．

step. 1　頻度情報または共起情報に基づいて重要単語を選択する．

step. 2　step1 で得られた単語を共起関係に基づき，いくつかのグループに分類する．この時，共起関係の指標は自己相互情報量 (PMI:Point-wise Mutual Information) を用い，予め設定された閾値以上のものを一つのグループとしてまとめる．

step. 3　step2 で得られたグループ内の単語と共起する単語を PMI を基に取得し，PMI が高いものを追加する．追加する単語数によって与える制約が変化するため，本研究では追加する単語数を 1~ 4 個で変化させ，トピックモデルの安定性を調べる．

4　実験

4.1　実験仕様

実験に用いる文書は，複数の文書からほぼ同一のトピックが抽出されるものが望ましいと考え，同じ話題の報告をしている複数の新聞記事を用いた．実験に用いた新聞記事は，アメリカ ABC News などより，2011 年 12 月 16 日の「野田総理による原発事故収束会見」に関する英字新聞 10 記事，2012 年 1 月 16 日の「イタリア豪華客船座礁事故」に関する英字新聞 24 記事，2012 年 1 月 17 日の「Wikipedia の SOPA への抗議」に関する英字新聞 25 記事，2012 年 1 月 17 日の「Yahoo! 共創業者の辞職」に関する英字新聞 18 記事である．

また，LDA-DF で用いるディリクレ分布のパラメータは，$\alpha = 0.1$，$\beta = 0.1$，$\eta = 100$ とし，推定方法は Collapsed Gibbs Sampling を用い，イレテーションは 50 回とした．トピック数 K = 10 とする．Hu ら [2] の研究では与えられた制約の単語に応じて，既存のトピックモデルで単語に割り当てられているトピックの一部を取り消し，潜在トピックの再推定を行っている．この取り消し対象となる単語の選び方に 4 つの方法を提案しており，その 4 つの中で文単位で取り消しを行った場合により良い結果が得られたと報告しているため，本研究でも再推定を行う際に取り消す単語は

文単位で行う．実験結果はパープレキシティ(式 (6)) の値を算出し，その値により事前知識を与える前と後でのモデルの安定性を比較する．ここで N は全文書長，w_{mn} は m 番目の文書の n 番目の単語，θ, ϕ はそれぞれ文書に対してトピックの生起確率，トピックに対して単語の生起確率を表す．

$$Perplexity(\boldsymbol{w}) = exp(-\frac{1}{N}\sum_{mn} log(\sum_{z} \theta_{mz}\phi_{zw_{mn}})) \quad (6)$$

4.2 実験結果

「野田総理による原発事故収束会見」の記事を例にすると頻度情報に基づく重要単語グループは {prime,minister,fukushima,reactor},{power,tokyo},{nuclear},{plant},{cold},{shutdown} となり，共起情報に基づく重要単語のグループは {cooling,contaminated,water},{stable,state,response},{worst,disaster},{site},{year} となる．これらのグループに PMI が高い単語を追加する．1 つ追加する場合の単語は頻度情報のグループにそれぞれ yoshihiko,electric,noda,march,reached,state となり，共起情報のグループにそれぞれ ton,tank,chernobyl,liquid,end となる．尚，本研究では，制約知識内で単語間の PMI を求め，その平均が高い制約知識から順に制約を与えることとした[1].

図 1 に「イタリア豪華客船座礁」の記事に関して，制約が 0 個の状態から制約を 1 つずつ与えていった場合のパープレキシティの変化のグラフを示す．このグラフから共起情報に基づいて構築した制約を与えた場合には，与える制約の個数を増やすにつれてパープレキシティが比較的低下する様子が見られ，制約を与えない通常の LDA の場合よりも安定したモデルとなることがわかった．重要単語に対して追加する語彙の個数は多くする必要はなく，1 つまたは 2 つ程度でパープレキシティを低下させられることがわかる．また，与える制約の個数は 3 個程度で最低値となり，その後もほぼ安定する様子が見られ，それ以上与える必要はないと考えられる．他の記事のグラフでも図 1 と同様のパープレキシティの変化が観測された．

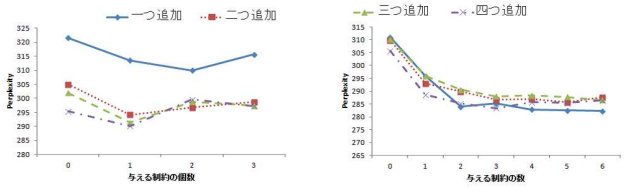

図 1: パープレキシティの変化 (左 : 頻度情報，右 : 共起情報)

5 考察

PMI に基づいて制約知識を構築した方が良かった理由としては，PMI は一文を単位として単語の共起関係を算出しているため，Andrzejewski ら [1] で用いた制約知識である Must Link と Cannot Link の両方を同時に反映した制約知識が作られたためだと考えられる．

表 1: 「野田総理の原発事故収束会見」記事から抽出されたトピックを代表する重要単語

topic	topic0	topic5	topic7	topic8
LDA	water plant contaminated remains	accident nuclear disaster chernobyl	year radiation plant area	cooling minister reactor prime
topic	topic0	topic5	topic7	topic8
制約付 LDA-DF	facility contaminated water cooling	nuclear disaster chernobyl worst	plant radiation year end	nuclear prime minister degree

また，共起情報に基づいて重要単語を決定した場合においても，制約の個数を増やしていった際にパープレキシティが上がっている場合も見られる．これは，潜在トピックが語の共起情報だけでは測れる訳ではないことを示していると考えられる．また，制約を与えてない LDA と制約を与えた LDA でのトピック分類の結果を見ると，制約を与えた LDA では制約で与えた単語群が同一トピックに入っており与えた制約知識を反映できていることがわかった．また表 1 にトピックに割り当てられた単語を示す．これは共起情報に基づいて重要単語を選択して構築した制約 {worst,disaster,chernobyl},{cooling,contaminated, water, ton},{year,end} などの制約を与えている．10 個のトピックのうち制約単語が現れているトピックのみ抜粋した．通常の LDA では topic8 に含まれている cooling が制約を与えた後では topic0 に入っているなど，制約知識を反映していることが分かる．

6 まとめ

従来の潜在的ディリクレ配分法ではユーザが期待していたものとは異なるトピック分類をされることがあり，それを改善するためにインタラクティブに制約を与えて再度トピック分類する研究 [2] などが行われてきたが，本研究では制約を対象文書から自動的に抽出しその制約を与えてトピック分類した際の結果を比較した．トピック分類された結果の比較方法は PMI やパープレキシティの他にもあるため，今後は他の指標を使って結果を比較するつもりである．また，今回，対象とした文書コーパスのサイズが小さかったため，今後はより大きなコーパスで調査をするつもりである．

参考文献

[1] Andrzejewski, D., Zhu, X. and Craven, M.: Incorporating domain knowledge into topic modeling via Dirichlet Forest priors, in *Proc. of*, ICML '09, pp. 25–32, New York, NY, USA (2009), ACM.

[2] Hu, Y., Boyd-Graber, J. and Satinoff, B.: Interactive topic modeling, in *Proc. of*, HLT '11, pp. 248–257, Stroudsburg, PA, USA (2011), Association for Computational Linguistics.

[3] Kobayashi, H., Wakaki, H., Yamasaki, T. and Suzuki, M.: Topic Models with Logical Constraints on Words, in *Proc. of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing* (2011).

[4] 鍛治伸裕, 喜連川優 : 語彙統語パターンにもとづく制約付き分布クラスタリング, 知識ベースシステム研究会, Vol. 79, pp. 61–66 (2007-12-03).

[1] 事前知識の組み合わせによっても PMI，パープレキシティの値は異なると思うが，今回はそこまでの精査はしていない．