4I1-R-9-3

共起情報に基づく事前知識を用いた潜在的トピック抽出の取り組み

An Approach to Extracting Latent Topics by Using Prior Knowledge based on Co-occurrence Information

立川華代　　小林一郎
Kayo Tatsukawa　　Ichiro Kobayashi

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
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

We have recently had many chances to treat huge documents. A lot of studies about extracting of latent topics in documents have been done, which is accomplished by Latent Dirichlet Allocation(LDA). That is not done by using superficial information. When we extract topics by LDA, sometimes it happens that the words we assume they are in the same topic is devided into different topics. Therefore David Andrzejewski and Yuening Hu proposed method that we select some words which should be in the same topic and incorporate the words into LDA as constrained knowledge. But the knowledge is constructed from subjective view of users in many cases and is not constructed from target documents automatically.

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．この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いる．LDA を使ってトピックを抽出する際に，本来，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．本研究では，与えられている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，Dirichlet Forest Prior を用いて，トピックとして分類される単語に制約を付与した．単語間の制約として，同様な確率値を持つ 2 つの単語が一緒にトピックに所属する制約として‘Must-Link’ を設定し，2 つの単語がいかなるトピックにおいても同時に大きな確率値にならないという制約として‘Cannot-Links’ を設定している．Hu ら [2] は，LDA を用いてトピック抽出された結果に対してインタラクティブに制約を与え再度トピック抽出を行う手法を提案している．また，Kobayashi ら [3] は，Andrzejewski ら [1] が採用した単語間の制約知識である Must Link と Cannot Link 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においては，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 [5] は，語彙統語パターンを用いて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．鍛冶らは単語分類における制約知識を対象とする文書からではなく，約 10 億文のコーパスから学習することによって獲得したが，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．

連絡先: 立川華代，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，〒 112-8610 東京都文京区大塚 2-1-1，03-5978-5708，tatsukawa.kayo@is.ocha.ac.jp

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．Dirichlet Forest の葉にあたる部分には各トピックにおいて出現する単語の出現確率が入り，全ての葉の出現確率の合計が 1 となる．

DF を用いた LDA（以下，LDA-DF）での文書生成過程では，まず通常の LDA と同様に，パラメータ α によって多項分布 θ を決め，これに従い 1 つのトピック Z を選択する．次にパラメータ β でおいて，定められた多項分布 ϕ において θ によって定められたトピック Z での確率分布に従い単語もくしくは制約を選択する．単語が選択された場合はその単語で文書を生成し，制約が選ばれた場合にはパラメータ η による多項分布 π において単語を選択することとなる．いま，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ると，LDA-DF は以下の式で表現される．

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

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

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

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

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

3.2 制約知識の構築

Newman ら [4] は，トピックの結束性に対する様々な評価指標について考察している．本研究においては，その指標の一つである，単語間の自己相互情報量（PMI:Point-wise Mutual Information）をトピック内の結束性を測る指標として採用し，'トピックらしさ' を測ることにする．

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し，以下に挙げる二つの手段を用いて選択する．

(i) 頻度情報に基づく重要単語の選択
同一内容の複数文書を対象にする場合，その全ての文書に共通して高頻度で出現する語は，その内容を表すために必要不可欠な語であると考えられる．そのため本研究では全文書に亘って現れる語を重要単語として選択することとする．

(ii) 共起情報に基づく重要単語の選択
本研究では PMI が高い単語どうしは，同一トピックに入る傾向が高いと仮定して制約を構築する．そこで，より多くの語と高い共起関係を持つ語を重要単語として選択する．これにより制約として与える単語群（事前知識に相当）間の PMI が高くなることが期待される．

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 PMI 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 を基にして取得し，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 個で変化させることとする．

4. 実験

4.1 実験仕様

トピック抽出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，イギリス BCC NEWS，カナダ CTV News など英語圏各国の主要新聞社や TV 会社のものであり，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 (文数は 212, 語彙数は 853)，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 (文数は 967, 語彙数は 2267)，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 (文数は 700, 語彙数は 1823)，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 (文数は 553, 語彙数は 1113),「パキスタンの雪崩」に関する英字新聞 22 記事 (文数は 403, 語彙数は 608),「北朝鮮ミサイル」に関する英字新聞 18 記事 (文数は 380, 語彙数は 936) を対象とする．

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 200 回とした．トピック数はパープレキシティに基づき対象コーパスに対して適切と思われるものを決定することもできるが，トピックの分類結果として得られる単語のグルーピングの適切さを同じ条件の下で直接判断したいことから，本研究においてはトピック数 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．このことから本研究でも，再推定を行う際に取り消す単語は文単位で行うこととする．

実験結果はパープレキシティの値を算出し，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．パープレキシティは式 (6) を用いて算出した．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

頻度情報および共起情報に基づく重要単語のグループとそのグループに追加される単語候補を表 1 に示す．表 1 中，抽出された重要単語を共起度が高いものを一つのグループに統合し，最終的に表 1 の上段に示す 6 つのグループを得た．この 6 つのグループそれぞれに単語を追加するが，1 つ追加する場合の追加単語を下段に示している．実際は下段の単語を上段のグループにそれぞれ追加することで制約知識を構築する．事前知識が 0 個の状態 (つまり通常の LDA と同じ) から一つずつ制約を加えていくことによって，トピック抽出精度が向上するかをパープレキシティを指標として確認する．

尚，本研究では，制約知識内の PMI が高いものから順に制約を与えることとした*1.

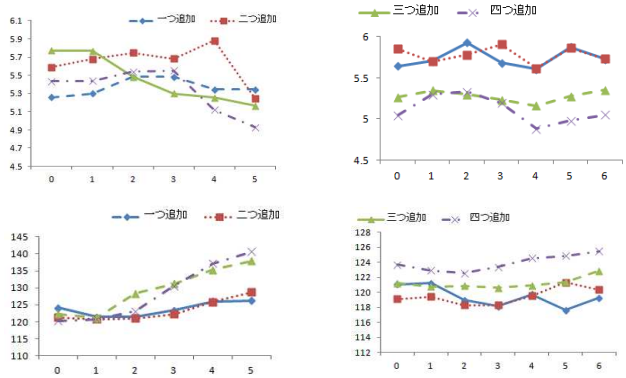

図 1: 「Yahoo!共創始者」(左：頻度情報，右：共起情報)

*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．

表 1: 頻度情報および共起情報に基づく重要単語グループと追加単語

種類/新聞記事		野田首相原発会見	イタリア豪華客船座礁	Wikipedia SOPA 抗議	Yahoo!共創始者辞職	パキスタン雪崩	北朝鮮ミサイル
頻度	重要単語によるグループ	{prime,minister, reactor,fukushima}, {power,tokyo}, {cold}, {nuclear},{plant}, {shutdown}.	{costa},{passenger}, {people}	{wikipedia},{online}, {piracy},{internet}	{yang},{board}, {yahoo},{company}, {thompson}	{pakistani}, {avalanche}, {soldier}, {troop}	{missile},{north}, {launch},{korea}
	追加する単語	yoshihiko,electric, reached,noda, march,state	appears,unaccounted, friday	wale,stop,protect,free	bostock,position, chairman,struggling scott	indian,army, saturday, remote	plan,long, rocket,japan
共起	重要単語によるグループ	{cooling, contaminated, water},{site}, {year} {stable,state, response}, {worst, disaster}	{disaster,caused,sea}, {aground,ran}, {gash}, {authority,safety}, {television}, {evacuation}	{medium,industry, group,tech,information, popular},{big}, {legislation}, {service},{community}	{private,pursuing, deal,shareholder, asian},{began}, {leaving}, {resignation},{chief}, {medium}	{civilian},{hour, survivor, entrance}, {snow},{located}, {metre},{morning, early,struggling, body}	{government}, {pyongyang}, {planned}, {report},{site}, {state,united,told ballistic,south}
	追加する単語	ton,liquid,end, tank,chernobyl	technical,late,side, trained,human, survivor	social,web,proposed, provider,wale	large,university, struggling,thompson, scott,trading	missing,entrance, headquarters,base ,area,make	image,peaceful, launching,news, allowed,chinese

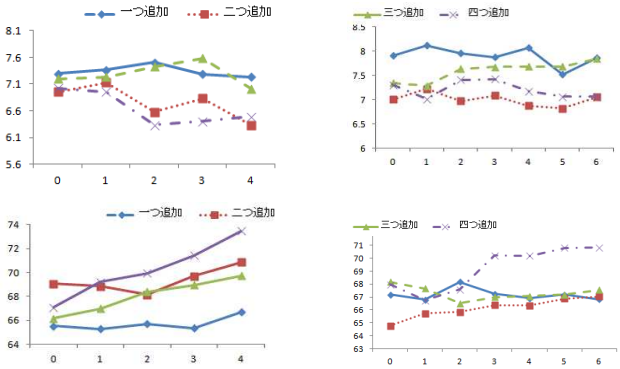

図 2: 「パキスタン雪崩」(左：頻度情報，右：共起情報)

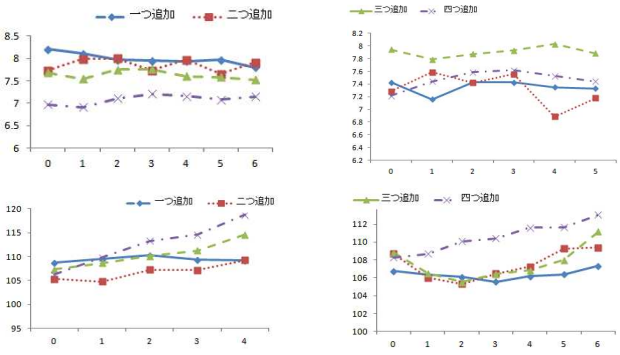

図 3: 「北朝鮮の衛星打上げ」(左：頻度情報，右：共起情報)

4.3 考察

図 1 から図 3 に「Yahoo!共創業者の辞職」,「パキスタンの雪崩」,「北朝鮮の人工衛星打上げ」の 3 つの記事に関して，頻度情報，共起情報についてそれぞれ与える制約数を増やしていったとき（横軸は与える事前知識の数）の PMI（図中上段）およびパープレキシティ（図中下段）の変化を示す．

ここで，“トピック”とはどのように表出されるかについて考えてみると，トピックとして纏められやすいものは，共起する概念が同一に分類されているものと考えられる．このことから，我々は語彙の共起情報を示す PMI とモデルとしての安定性を示すパープレキシティを用いて，“良いトピック分類”とは「PMI が高くパープレキシティが低いトピック分類」であるという仮説を立てる．

PMI とパープレキシティの関係を捉えるため，事前知識に共起する単語を 1 から 4 つまでを増やした際の相関係数を示す[*2]．図 4 に頻度情報に基づいて事前知識を作った際の PMI とパープレキシティの相関関係を，図 5 に共起情報に基づいて事前知識を作った際の PMI とパープレキシティの相関関係をそれぞれ示す．

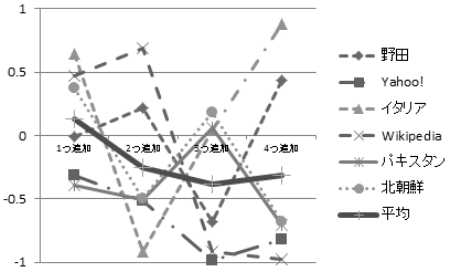

図 4: PMI とパープレキシティの相関関係（頻度・事前知識）

図 4 において，文書によって相関係数の値の取り方に差異が生じているが，全体の平均を見ると，頻度情報に基づいて構築した制約知識は，PMI とパープレキシティの間に値は小さいが負の相関関係が存在し，単語を追加することにより，値が

*2 この相関係数は事前知識の個数の追加をひとつずつ増やしていった際に共起する単語を追加した場合のものである．

表 2: 「北朝鮮の人工衛星打上げ」記事から抽出されたトピックを代表する重要単語（上位 10 個）

topic	topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
LDA	missile north rocket nuclear technology korea expert satellite long range	missile launch jang ballistic jin test myong satellite facility program	north ri site taongchang rocket stage launch south reporter west	north korea satellite launch space preparation station foreign security council	north launch kim korea april il sung satellite birth founder	launch north rocket long range stage warning korean international position	south korean nuclear news test satellite government party report tunnel	weather send military back image forecast survey people risk suspended	kim jong power satellite meeting leader december told timing authority	korea north test missile state nuclear japan rocket united part
topic	topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
制約付 LDA-DF	missile north rocket long expert range state united technology plan	jin people myong jang space show develop facility country taraget	north rocket range long japan plan kwangmyongson satellite orbit urged	ri north tongchang south station rocket preparation japan reproter long	launch satellite il sung month april founder birth north kim	space stage security council liftoff international fueling position send scheduled	south test korean news government party image sunday report agency	send korean weather military forecast back survey conduct resource aid	jong power meeting chine december high leader father minister chinese	test korea nuclear north week foreign kim needed range long

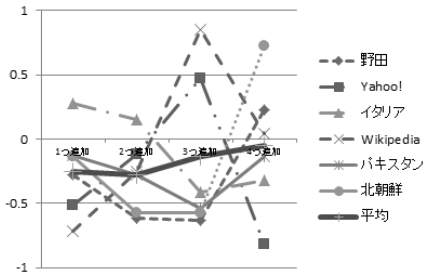

図 5: PMI とパープレキシティの相関関係（共起・事前知識）

増加する傾向にあることがわかる．PMI とパープレキシティ間における負の相関は，我々が立てた仮説である「PMI が高くパープレキシティが低いトピック分類」を支持するものであり，頻度情報から重要語を抽出し，共起する単語を追加することで構築する制約知識が良いトピック分類を実現するのではないかと考えられる．一方，図 5 においても文書によって相関係数の値の取り方に差異が生じており，全体の平均をとっても相関係数の値が小さく，単語を追加しても負の相関が現れるとは言い難い．

上記のことから，トピック分類に有効となる制約知識は，文書から頻度情報に基づき重要語を抽出した後に，それらの語彙から制約知識の核となるグループを構築し，そのグループに共起する単語を追加することから構築すれば良いと考えられる．

表 2 に，「北朝鮮の人工衛星打上げ」の記事に関して，制約を与えていない通常の LDA(上段) と制約（制約知識 5 つのそれぞれに対し PMI の高い単語を 4 つ追加したもの）を与えた LDA-DF(下段) でのトピック分類の結果を示した．制約は頻度情報に基づいて構成された，{missile,technology,state,expert,united},{north,long,rocket, plan,japan},{launch,il,sung,month,satellite},{korea,kim, foreign,week,nuclear} の計 4 つの制約を与えた．これにより，実線で示される制約単語が上段では Topic0，1，9 に分かれていたものが下段では Topic0 に集結し，一点鎖線で示される制約単語が上段では Topic7，9 以外に現れていたものが下段では Topic4 に集結し，点線で示される制約単語が上段では Topic0，3，4，6，8，9 に現れていたものが Topic9 に集結していることがわかる．

5. おわりに

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．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，今後は他の指標の下でより大きなコーパスを使い，結果を比較するつもりである．また，制約の構築方法自体も単に PMI の高いものどうしを制約として与えるだけでなく，より良い制約となるものを構築する手段を検討するつもりである．

参考文献

[1] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proc. of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 25–32, New York, NY, USA, 2009. ACM.

[2] Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. Interactive topic modeling. In *Proc. of the 49th ACL-HLT - Volume 1*, pp. 248–257, 2011.

[3] Hayato Kobayashi, Hiromi Wakaki, Tomohiro Yamasaki, and Masaru Suzuki. Topic models with logical constraints on words. In *Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*, 2011.

[4] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *HLT: The 2010 North American Chapter of the ACL*, pp. 100–108, 2010.

[5] 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.