6C-4

# トピックモデルで生成したトピックへ付与するラベルの自動生成

慶留間 諒大† 當間 愛晃‡ 赤嶺 有平‡ 山田 孝治‡ 遠藤 聡志‡
†琉球大学理工学研究科情報工学専攻 琉球大学工学部情報工学科‡

## 1 はじめに

文書が持つ潜在的なトピックを抽出するための手法として、pLSI や LDA[1] などの手法がある。これらはトピックモデルと呼ばれており、情報検索 [3] や複数文書要約 [4] など、様々な分野で応用されている。

トピックモデルによって抽出されたトピックは、トピックに関連する単語とそれぞれの単語がそのトピックにおいて生起する確率の集合で表され、どのようなトピックであるかはこれらの情報を基に判断する必要がある。

トピックを解釈する上で最も標準的な方法は生起確率の高い単語からトピックの内容を推測する事だろう。しかし、生起確率が高い語が必ずしもトピックの解釈に役立つとは限らない。その例として、簡単なトピック抽出の結果の例を表 1 に示す。表の枠内左側がそのトピックに属する単語であり、右側がそのトピックにおける単語の生起確率である。ここで示されている単語は、そのトピックにおける生起確率が上位 5 件以内のものである。トピック 1 は関連する単語からユーザー解析に関したトピックであると推測する事が出来るが、トピック 2 に関してはトピックに関連する単語からどのような内容であるかを推測する事は難しい事が分かる。そのため、トピックの内容を判断するにはあるトピックにおける単語の生起確率のみでは不十分である。

表 1: トピック抽出の結果の例

| トピック 1 | |
|---|---|
| ユーザー | 0.072554 |
| 的 | 0.066291 |
| 解析 | 0.054286 |
| 発言 | 0.030277 |
| 批判 | 0.026624 |

| トピック 2 | |
|---|---|
| 評価 | 0.091736 |
| 文節 | 0.060658 |
| 条件 | 0.033088 |
| 表現 | 0.029078 |
| 素性 | 0.026572 |

また、トピックの解釈が解釈する人の主観に依存していたり、トピック抽出を行った文書群に関する知識が不十分である場合にどういうトピックであるかを上手く解釈できないなどの問題があり、このような問題を解決するために抽出したトピックに対して自動でラベリングを行うという研究が進められている。

‡A basic study about automatic label generation on topic models.
†Ryouta KERUMA ‡Naruaki TOMA ‡Yuhei AKAMINE ‡Koji YAMADA ‡Satoshi ENDO
†The Graduate School of Engineering and Science, University of the Ryukyus
‡The Department of Information Engineering, Faculty of Engineering, University of the Ryukyus

## 2 ラベル生成における諸問題

### 2.1 関連研究

Mei ら [2] は、Chunking、または N-gram Testing によって文書群から候補ラベルを生成し、生成されたラベルを KL-divergence ベースの手法を用いてランク付けして候補ラベルを選択するという手法を提案した。Chen ら [5] は、LSA(潜在的意味解析) を用いて抽出したトピックに対して相互情報量と n-gram を用いてそのトピックの内容を表すフレーズを生成するという研究を行った。嘉村ら [6] は、トピックの客観的な解釈を可能とする単語をトピック表現語と呼び、トピック表現語の特徴量を提案し、トピック表現語を抽出する試みを行った。黒田ら [7] は、Web 文書をクラスタリングして生成されたクラスタに対して、クラスタ間の差異を考慮して重要語を抽出し、タイトルパタンを適用する事でクラスタの内容を表すタイトルを生成する研究を行った。

### 2.2 ラベルの種類と特性

ラベル生成の研究では、ラベルは単語や n-gram で生成したフレーズといったようにいくつかの種類に分かれており、どれを使うのがラベルとしてふさわしいかを検討する必要が有る。

まず、単語についてだが、一つの単語をラベルとして利用する場合、解釈の手がかりにはなるが具体的にどのような内容がそのトピックに含まれているかを理解するには不十分になりがちである。例えば文書要約関連のものと思われるトピックが抽出された時、そのトピックの内容を一つの単語で表すのならば「要約」という単語がふさわしいだろう。しかしこの場合、単一文書を対象とした文書要約なのか、もしくは複数文書を対象とした要約なのかという事までは分からない。そのため、もう少し多くの情報を含んだラベルを生成する必要が有るだろう。

次に、n-gram であるが、こちらは Mei ら [2] の研究や、他のラベル生成の研究においても多く用いられており、一定の成果を挙げている。しかし、これらの研究では英語を対象にしており、日本語では良いラベルが生成されるとは限らないため検証する必要がある。

## 3 研究内容

今回、我々は Chen ら [5] の研究において紹介されていた手法を用いてトリグラムでフレーズを生成する実験を行った。

Chen らの研究では LSA でトピックの抽出を行ったが、我々は LDA でトピックの抽出を行った。その理由は、LSA よりも LDA の方が一般に広く使われている

ためである。また、Chen らの研究で提案された手法は LDA の場合でも特別な変更を加える必要はないという利点もある。

### 3.1 ラベル生成

Chen らの研究では、対象文書群の各文書から、n-gram でフレーズを抽出し、相互情報量が閾値を超えた者をラベル候補として扱う。相互情報量の例として、バイグラムの場合の式を以下に示す。

$$MI(x,y) = log\frac{N_{xy}/N_b}{(N_x/N)(N_yN)} \quad (1)$$

x と y はそれぞれバイグラム中の単語である。$N_{xy}$ は、バイグラム xy の出現回数であり、$N_b$ は全バイグラムの出現回数の総和である。$N_x$,$N_y$ はそれぞれ単語 x と単語 y の出現回数であり、N は全単語の出現回数である。

### 3.2 ラベル評価

ラベルの評価は LSA-weighted frequency(LSAF) という手法を用いる。計算式は以下のようになる。

$$LSAF(xy) = |\sum_{i=1}^{d}(\alpha_i(N_{i,x} + N_{i,y})/2)| \quad (2)$$

d は対象文書群の文書数である。$\alpha_i$ は、文書 i におけるトピックの重みを表すベクトルである。$N_{i,x}$,$N_{i,y}$ は、それぞれ文書 i において単語 x と単語 y が出現する回数である。

## 4 実験

### 4.1 実験設定

今回対象とした文書は NLP2013 の要約セッションと評判・感情分析セッションで発表された 10 部の論文であり、トピック数は 8 とした。また、Mecab を用いて形態素解析を行って抽出した名詞のみでラベルを生成している。

### 4.2 結果と考察

実験結果を表 2 に示す。

表 2: トリグラムでラベルを生成した例

| トピック | ラベル |
|---|---|
| topic 0 | 感情 極性 推定 |
| topic 1 | 批判 的 ユーザー |
| topic 2 | 評価 条件 文節 |
| topic 3 | 解析 見出し 文 |
| topic 4 | 複数 文書 要約 |
| topic 5 | 部分 木 集合 |
| topic 6 | 情緒 名 情緒 |
| topic 7 | 意見 ツイート 抽出 |

表 2 の結果を観察すると、トピック 0 の「感情 極性 推定」、トピック 4 の「複数 文書 要約」、トピック 7 の「意見 ツイート 抽出」は単一の単語の場合と比べて理解しやすいものとなっている事が分かる。他方、トピック 2 の「評価 条件 文節」やトピック 6 の「情緒 名 情緒」は具体的にどのようなトピックが含まれているかを推定する事は難しい。この問題はラベルを生成する際に利用している単語の評価法を変えたり、もう少し幅広い品詞を利用するなどの工夫により改善するかもしれないが、文書中にトピックを解釈する上で役に立つような n-gram が存在しているかどうかという点も問題になるだろう。

## 5 今後の予定

現段階では前節で紹介した n-gram でのフレーズ生成の実験を行った所に留まっている。前節での実験を踏まえて、まず、適用する品詞の範囲を広げる事でより有用なラベルが生成されないかを確認したいと思っている。また、トピックモデルを適用した結果として得られるのはトピックを構成する単語とその生起確率だけでなく、それぞれの文書におけるトピックの割合も得られる。このある文書が持つトピックの割合を利用して、あるトピックの割合が大きい文書から、そのトピックのラベルにふさわしい文などの情報が抽出できないかと考えている。

## 参考文献

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

[2] Q. Mei,X. Shen,and C. Zhai,“Automatic labeling of multinomial topic models”,In SIGKDD,pp.490-493,2007

[3] S. Wei,W.B.Croft,“LDA-based document models for ad-hoc retrieval” In SIGIR ’06,pp.178-185 (2006)

[4] Aria Haghighi, Lucy Vanderwende,“Exploring Content Models for Multi-Document Summarization”,In HLT:NAACL 2009,pp.362-370 (2009)

[5] J. Chen, J. Yan, B. Zhang, Q. Yang, and Z. Chen,“Diverse topic phrase extraction through latent semantic analysis”,In Proceedings of ICDM ’06,pp.834-838 (2006)

[6] 嘉村 亘太, 黄 宏軒, 川越 恭二,“トピックモデルにおけるトピック表現語導出”,DEIM フォーラム 2012,(2012)

[7] 黒田 知宏, 内海 彰,“文書クラスタリングにおけるクラスタタイトルの自動生成”, 情報科学技術フォーラム講演論文集 7(2),pp.279-281 (2008)