4Q-08

可変長N-gramに基づいたトピックへのラベル選択の検証

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

1.1 研究背景

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単語の確率分布で表され、どのようなトピックであるかはこれの情報を基に判断することになる。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、この方法ではトピックの解釈が解釈する人の主観に依存してしまうという問題がある。Lauら[5]の研究では上位10単語の中からトピックのラベルとしてふさわしい単語を被験者10名に選択させたところ、最も選択された単語でも被験者の5割程度しか選択されなかったという報告がある。また、単語からトピックを解釈するという方法は、文書群で扱われているトピックに関する知識が不十分である場合にトピックを上手く解釈できないといった問題も存在する。近年では単語単位ではなくフレーズ単位でトピックを抽出する[9]などの方法も提案されているが、いずれにしてもトピックを抽出した後にその内容を解釈するという過程は存在しており、トピックの内容を決定的に表すラベルを付与する研究は必要であると考えられる。

これまでのラベル付与の研究ではN-gramからラベルを選択する研究[2]やトピックモデルで得られた単語分布の上位単語からラベルとしてふさわしい単語を選択するといった研究[5]が行われている。N-gramからラベルを選択する研究ではこの段階では固定長のN-gramを取り扱っているため、ラベル候補が限定されているという問題がある。また、単語からラベルを選択する方法は既に述べたように解釈者の主観による違いが大きくなるという問題がある。本稿ではこれらの問題を踏まえて固定長ではなく可変長のN-gramからトピックのラベルを選択する事を目指す。固定長のN-gramよりも広い範囲のN-gramを取り扱う事で、より尤もらしいラベルが得られる事が期待される。

1.2 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ベルを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。Meiらの評価手法は、同じ単語が何度も出るようなノイズに近いラベルの評価が抑制されるという利点があり、本研究ではこの評価方法を採用した。

2 研究内容

まず、対象とする文書群の中から可変長のN-gramを抽出する。次に、文書群からトピックの抽出を行い、得られたトピックに対してN-gramの中からラベルとしてふさわしいN-gramを選択した。ラベルの評価方法には上述の通りMeiら[2]の方法を利用した。

2.1 可変長N-gram抽出

可変長N-gramの抽出には持橋ら[7]が提案したVPYLMを利用した。VPYLMは与えられた文脈から確率的にその長さを決定する事で可変長のN-gramを抽出する事を実現しており、理論的には無限長のn-gram抽出を行う事が可能である。VPYLMではN-gramの最大長を固定した場合においても余分な計算が行われにくいために高速な計算ができるという利点がある。

この方法で抽出したN-gramは抽出したN-gramとその生起確率のペアで表されるが、確率がスムージングされるためにイテレーションが進む毎に計算されなくなった不要なN-gramの生起確率も計算できてしまう。また、生起確率は全N-gramに対しての確率ではなく、与えられた文脈からそのN-gramが得られる確率であるため、Meiらの評価法を利用するには多少手を加える必要がある。そこで力押しな方法ではあるが、今回はN-gramを抽出した後に文書群から各N-gramが得られる回数を直接計算した。また、今回の実験では最大長を8として実験を行った。これは最大長を大きくする事で不要なN-gramが生成されるのを抑制するためである。

2.2 トピック抽出

トピックの抽出にはTehら[8]が提案したHDP(Hierarchical Dirichlet Process)を利用した。HDPはまず各文書内の単語をクラスタリングし、その後で全文書のクラスタをクラスタリングする事で文書をトピック毎に分ける。HDPの利点は自動でトピック数も推定するために事前にLDAなどのようにトピック数を確定させる必要が無い点である。

---

‡A Simulation of Label Selection for Topics using Variable Length N-gram
†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

ラベル	重要文抽出による要約	製品開発における不具合事例の利用	発信者情報を抽出する
単語	文	語	発信
	要約	製品	情報
	重要	平易	者
	問題	不具合	抽出
	生成	化	日時

表 1: 可変長 N-gram ラベルと上位単語リスト

2.3 ラベル評価

Mei ら [2] の評価方法では与えられた文書群 C からラベル l が得られる確率 $p(l|C)$ と文書群 C から単語 w が得られる確率 $p(w|C)$ の PMI を利用する。計算式は以下の (1) で表される。

$$\begin{aligned} & Score(l,\theta) \\ &= \sum_{w} p(w|\theta) log \frac{p(w,l|C)}{p(w|C)p(l|C)} - D(\theta||C) + Bias(l,C) \\ &= \sum_{w} p(w|\theta) PMI(w,l|C) - D(\theta||C) + Bias(l,C) \quad (1) \end{aligned}$$

右辺第 2 項は文書群とトピックの距離であり、こちらはどのラベルでも共通なので実際には計算しない。また、第 3 項のバイアス項も基本的には 0 と仮定するので第 1 項のみを評価に利用する。

3 実験

3.1 実験設定

今回対象とした文書は自然言語処理学会の年次大会発表論文集 2010 年度から 2013 年度の 4 年分であり、この文書群を Mecab で形態素解析したものを実験に使用した。可変長 N-gram の抽出の際にはストップワードや低頻度単語を除去すると適切な N-gram が取り出せないため頻度 10 以下の単語を一律に同じ文字に置き換えた。また、トピック抽出の際には頻度 10 以下の単語とストップワードの除去を行った。

3.2 結果

表 1 にトピックモデルで得られたあるトピックの単語分布の上位 5 単語とスコアが最も高かったラベルを示す。可変長 N-gram を利用する事で単語よりもある程度トピックの具体的な内容を示すラベルを得る事ができた。しかし「コーパスの」などの中途半端な所で切れてしまっているラベルや、「機械翻訳システムが対訳言語」といった少し冗長な N-gram がラベルとして選択されてしまうといった事例も見られた。

4 今後の課題

可変長 N-gram を利用する事である程度の成果は得られたが冗長性や半端に切れた N-gram をどう扱うかが問題が残った。これは情報量の和でスコアリングをしているので冗長なラベルの方がより情報が多いという事でスコアが上がったと考えられる。今後は余分な情報を含んでいる時の取り扱いをどのようにするかが課題である。

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] Lau, Jey Han and Newman, David and Karimi, Sarvnaz and Baldwin, Timothy,“Best Topic Word Selection for Topic Labeling”,In Proceedings of the 23rd International Conference on Computational Linguistics: Posters,pp.605-613 (2010)

[6] Lau, Jey Han and Grieser, Karl and Newman, David and Baldwin, Timothy,“Automatic Labelling of Topic Models”,In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1,pp.1536-1545 (2011)

[7] 持橋大地, 隅田英一郎 “階層 Pitman-Yor 過程を用いた可変長 n-gram 言語モデル”, 情報処理学会論文誌,Vol.48,No.12,pp4023-4032(2007)

[8] Teh Y. W., Jordan M. I., Beal M. J., Blei D. M. “Hierarchical Dirichlet Process”,In Journal of the American Statistical Association 101: pp1566—1581(2006)

[9] 若林啓 “階層型 HMM に基づくフレーズ生成トピックモデルの提案”,DEIM フォーラム 2014