E-007

トピックモデルを用いた議会議事録分析

Analysis of Congressional Proceedings with Topic Models

小島智樹[1]　　八槇博史[1]

Tomoki Kojima　Hirofumi Yamaki

1　はじめに

国の動向を決定する議会の議事録を利用し,分析することで国の動向の予測を行うシステムの作成を行う.政治,経済について議論している議会議事録には防衛や医療などの様々な分野の用語が含まれている.この議事録を機械学習で利用できるようにベクトルに変換する際,推論ベースでは利用される用語の多様性から議事録の特徴を抽出することが難しい.そこで,カウントベースであるトピックモデルを用いて単語ごとにどの分野の用語であるかを判別し議事録の分析を行う.分析したデータを利用し,議事録同士の派生関係を求め収集する.このデータを議会の行動と結果のデータセットとし,国の動向を予測するシステムの作成を目指す.

2　トピックモデル

確率的生成モデルの一つ.単語の共起性をモデルで表すことを目指したもの.それ以前のトピックを推定するモデル(混合ユニグラムモデル)では,一つの文書が一つのトピックを持つと仮定し計算を行っていた.対してトピックモデルは,文書が複数の潜在的なトピックから生成されると仮定するモデル.また,文書内の各単語はトピックが持つ確率分布に従って出現すると仮定する.これにより,「オリンピックの経済効果」など,「スポーツ」,「経済」などの複合的なトピックを含む文書の分類を可能とする.

2.1 Latent Dirichlet Allocation (LDA)

トピックモデルの一つであり,多くのトピックモデルの基礎となるモデル.文書中の単語の順序は無視し,Bag of Words (BoW)表現と呼ばれる単語と出現頻度のペアの集合をモデル化する.BoW 表現は単語が共起している現象を表している.LDA では特に一つの文書には複数の潜在トピックが存在すると仮定し,そのトピックの分布を離散分布としてモデル化する.

潜在意味解析(LSI)という特異値分解を用いて潜在トピックを推定する方法を改良し,確率モデルとして取り扱えるようにしたものを確率的潜在変数モデル(PLSI)と呼ぶ.この PLSI が新しい文書の確率が未定義となり計算が困難であったという点に着目し,ベイズ化したモデルが LDA となる.

本研究では,この LDA を用いて議事録からトピックを推定し,分類を行う.

2.2 Coherence

確率モデルの汎化能力を測る指標.トピック中の単語の性質に関する評価にも利用される.数値が高いほど性能がいいものとなる.本研究では確率モデルの汎化能力を評価するために利用する.

2.3 Perplexity

確率モデルの性能評価の尺度.Perplexity の値は選択肢の数を示している.低い値をとるほど高い精度で予測できる確率モデルとなる.

3　利用データ

本研究ではインターネット上に公開されているイギリス議会議事録のデータを利用し,分析することで研究を行っていく.利用する議会議事録

[1] 東京電機大学大学大学院システムデザイン工学研究科
Graduate School of System Design Engineering, Tokyo Denki University

は,1980 年から 2005 年までの議事録で,その中から 1111 件のデータを利用する.

4　実験（研究）結果

Coherence，Perplexity の 2 つの評価指標を基に,データセットのトピック数を決定していく.それぞれの計算結果を表示したものは以下の通りとなる.

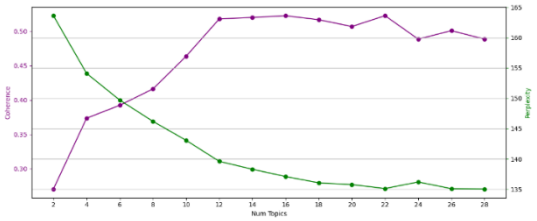

図 1Perplexity，Coherence の推移

分類するトピックの数が増えるほど精度が高くなり,安定していったが,トピック数 22 個以降は,Perplexity は緩やかに低下しているが,Coherence は低下していく結果となった.そのため,22 このトピックに分類することする.LDA を用いて分類をし,語句をレマタイズとステミングを行い,単語を原型に変換し,出現回数が 100 回以下のもの,出現率が 10%以上の単語の除去を行った.トピックごとの利用頻度の高い単語は図 2,3 の通りとなった.

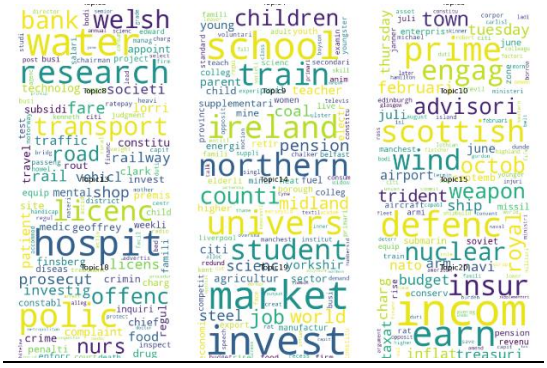

図 2 トピックごとの頻出単語 1

図 3 トピックごとの頻出単語 2

トピックの中身の単語を見ると学校に関すること,科学に関すること,医療に関すること,貿易に関することに分かれている.

5　おわりに

今回議事録の派生関係を求めるうえで必要となる議事録の分析を行った.議会の議事録の内容は多くの分野に触れており,推論ベースの手法による文書のベクトル化を利用すると特徴が似てしまい分類が難しく,カウントベースの手法を用いたが,想定よりも高い精度で分類ができていた.

しかし,トピックごとの単語が人間にもわかりやすい分類となったが,定量的に何に関するトピックであるかを評価する必要があること,派生関係を求める際に議事録それぞれがどのトピックに関する話題に触れているか,それが議事録の中でどれほど重要な話題となるかといった情報も評価できる指標の必要があることが分かった.

参考文献

[1] 佐藤一誠，奥村学(監修)，“トピックモデルによる統計的潜在意味解析”，コロナ社，(2015-4)

[2] 岩田具治，“トピックモデル”，講談社，(2015-4)