Vol.2015-NL-223 No.4
2015/9/27

政策形成過程における 文書自動比較システムに関する応用研究

加藤 大暁[1]　木下 貴文[1]　横澤 誠[1]

概要：自然言語処理を用いた政策文書分析に関する研究は，政党のマニフェストや政治家のスピーチ文から政治的立場の分析などが行なわれている．現在，政策形成過程における国際会議では，各国の主張や法制度の複雑化された関係を人の手によって比較表の形式にまとめ，視認性を高めた上で最終的な判断を決定する事が行なわれている．本研究は，政策形成過程における比較分析を補助する文書比較システムの開発を行なっており，複数文書間の同一トピックににおける内容の差異を抽出する事を目指す．本稿では，文書比較の最初の試みとして簡単な政策に関する文書に対して特徴抽出し，トピックごとのクラスタリングを行ない，評価する．

1. はじめに

自然言語処理を用いた政策文書分析に関する研究は，政党のマニフェストや政治家のスピーチ文から政治的立場を分析などが行なわれている．国際公共政策に関する会議では，各国が様々なトピックに関する主張を記述した寄与文書が事前に提出される．各国によって主張や法制度が異なるため，各国間の立場が非常に複雑化された上で，意思決定や合意形成を行なわなければならない．それらの関係の視認性を高めるために，人の手によって比較表の形式にまとめることが行なわれている．しかし，数多くの主体が提出する莫大な文書を限られた時間で分析する必要があり，非常にコストがかかっている．

本研究では，政策形成過程における比較分析を補助する文書比較システムの開発を行なっており，複数文書間の同一トピックににおける内容の差異を抽出する事を目指す．まず，対象の文書から同一トピックに関する文章を抽出し，比較表の形式で文書を提示する．文書中のトピックを比較項目とし，各国の文書に関して同一トピックに関してまとめる．対象とする政策文書は，パラグラフ単位でトピックが決まり，同一トピックに関するパラグラフを構成する単語は類似しているの考えられる．そこで，本問題を対象文書中の類似パラグラフに対するクラスタリング問題とみなす事ができる．

本稿では，比較表作成の最初の試みとして，簡単な政策文書に対して既存の手法を用い，前処理，特徴抽出，クラスタリングを行ない結果を評価した．

[1] 京都大学情報学研究科
Graduate School of Informatics, Kyoto University

2. 関連研究

一般的な文章に対する類似文書抽出は，剽窃の発見 [1]，質問応答タスク [2] などで行なわれている．Bär[1] らは，剽窃の発見を目的とし内容類似度，構造類似度，形式類似度の 3 次元の特徴を用いた文章の類似度を提案している．内容類似度とは，TF-IDF 法などを用いて文書中に存在する単語の出現類似度を計算したものであり，構造類似度とは，単語の順序などの文法的類似度を n-gram などで計算したものであり，形式類似度とは，語彙の豊富さを異語率（type-token ratio）などによって計算したものである．Celikyilmaz ら [2] は質問応答（Question Answering）タスクにおける質問文と回答候補の Ranking に，自然言語処理において用いられるトピックモデルである LDA，hLDA（Hierarchical LDA）を類似度計算に用いている．

専門的な文書を対象として類似文章の抽出を行なう研究として，特許調査 [3]，診断文書検索 [4] が挙げられる．目黒ら [3] は「F ターム」という日本の特許文献に人手で付与されている分類記号に含まれる「F ターム観点」の付与されやすさを数値化した F ターム概念ベクトルを作成し，特許文献間の類似度を計算している．また，岡本ら [4] は診療文書の定型性である，”Observation”，”Diagnosis”，”Treatment” のサイクルの着目して類似度を計算している．各文章がどのサイクルに当たるかを人の手によってラベル付けされた教師データを基に学習を行ない，対象文書を TF-IDF 法によってラベル付けする．対象となる診断文書と同一サイクルの文中の逆文書頻度（Inverse Document Frequency）

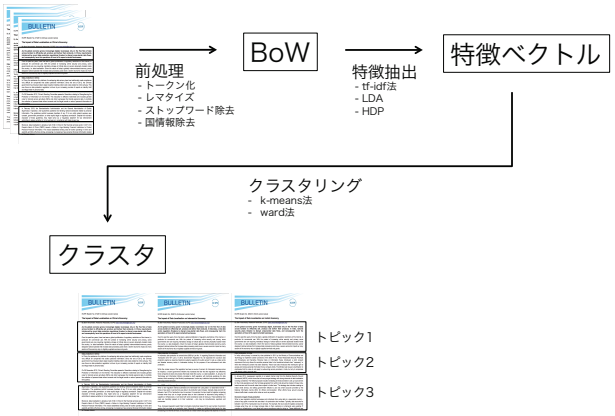

図 1 フローチャート

から類似度を計算する事で，類似診断書の検索を行なっている．これらの研究はそれぞれの分野の特徴的な情報をラベルとして用いて分析しているが，本研究で対象とする文書では分類情報が付与されていないため，用いる事ができない．

政策文書に関する言語処理の研究は，マニュフェストの分析や政治家のスピーチの分析などが行なわれている．従来の研究 [5][6] では各政党や政治家の全般的な立場の分析だったが，Zirn ら [7] はドイツの立法機関であるドイツ連邦議会の発言内容から，それぞれの政治家がどのトピックにどのような立場を取っているかを分析している．会議の告知文書から抽出した候補となる単語を，専門家に整理してもらうことで，ラベルを決める．すべての政党の公表するマニュフェストに対して Labeled-LDA モデル [8] によってラベル付きの文書を学習し，実際の会議の発言のトピックを推定する．さらに，抽出した各トピックに関する発言から，その政治家が政党のマニフェストに対してどれだけ異なった立場を取っているかをの分析を行なう．各政党のマニュフェストに対して Labeled-LDA モデルによって学習し，マニュフェストのトピック分布と政治家の発言から得られたトピック分布との差異を計算している．

3. システム概要

本章では，入力された政策文書のクラスタリングを行なうまでの各ステップとそこで手法について説明を行なう．本稿で実行した処理は，入力したデータに対する前処理，各パラグラフの比較のためのベクトル空間モデル の特徴抽出，それぞれの特徴ベクトルに対するクラスタリングの 3 ステップに分かれており，図 1 で示す．以下では各ステップについて詳しく説明する．

3.1 前処理

前処理は以下の 3 ステップに分かれている．

トークン化 パラグラフ単位ですべての文書を読み込み，文章のトークン化を行なう．トークン化とは，文字列をトークンと呼ばれる文章を構成する最小単位に分割する事であり，英語を対象とする場合は，ほとんどは空白文字によって分割する事が可能である．

レマタイズ 各トークンを WordNet レマタイザによって見出し語化を行なう．WordNet レマタイザでは，対象のとなる単語の見出し語が辞書に存在する場合に見出し語に変換する．その結果，”words”と”word”などの単語を同一単語として扱うことが可能になる．

ストップワードの除去 ストップワードの削除を行なう．ストップワードとは，”a”や”the”一般的にどんな文書にも頻出する語であり，情報量が少ないため対象外とする．

国情報の除去 複数の国にを超えてクラスタリングするために，国名に関する単語も削除する．国名が頻出した際に，その国の文書中の異なるトピックに関するパラグラフを国名によって同一トピックとクラスタリングしてする可能性があるからである．

python の自然言語処理ライブラリである nltk を用いて行なった．

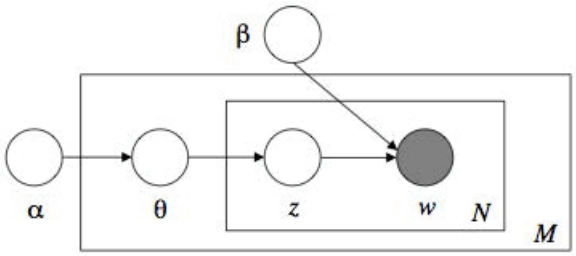

図 2 LDA のグラフィカルモデル

3.2 特徴抽出

前処理を行なった文書集合をパラグラフ単位での Bag-of-Words（BoW）表現に基づくベクトル表現に変換し，BoW を以下の手法で特徴ベクトルに変換する．python のトピックモデル用ライブラリである gensim を用いて実装した．

tf-idf 法 tf-idf 法とは単語の出現頻度とその単語の珍しさによってと BoW によるベクトル表現 に重み付けを行なったものである．文書におけるある単語の出現頻度を TF(Term Frequency)，ある単語が文書に出現する頻度の逆数を idf（Inverse Document Frequency)，tf と idf をかけた tf-idf を特徴量として利用する．

LDA LDA[9] は対象文書からトピックの抽出を行なうトピックモデルの代表的手法である．トピックモデルとは，1 つの文書を複数のトピックの混合分布で表現する生成モデルであり，以下の生成過程でパラメータを推定し，トピックを抽出する．また，その際のグラフィカルモデルは図 2 で表現される．

(1) 単語数 N をポアソン分布によって選択する

$$N \sim Poason(\zeta)$$

(2) トピックの確率分布 θ を Direchlet 分布に基づき選択する

$$\Theta \sim Dir(\alpha)$$

(3) 各 N 個の単語 w_n に対して以下を繰り返す.

(a) 多項分布 $Mult(\Theta)$ に基づきトピック z_n を選択する

$$z_n \sim Mult(\Theta)$$

(b) トピック z_n に対する単語の確率分布に基づき単語 w_n を選択する

$$w_n \sim P(w_n|, z_n, \beta)$$

以上の過程を M 回繰り返し，文書集合 D を生成する．パラメータ α と β が与えられたとき，文書集合 D の生成確率は以下のように表される．

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_(dn)} p(z_d n|\theta_d) p(w_d n|z_d n, \beta)) \right) d\theta_d$$

パラメータ α と $beta$ の学習には崩壊型ギブスサンプリングを用いた.

トピックの混合分布 θ_d をベクトル空間に表し，パラグラフ単位で以下のような特徴ベクトルで表す．$F_p = (\theta_{p1}, \theta_{p2}, ..., \theta_{pK})$，$\theta_{pk} = p(k|\theta_d)$ は，パラグラフ p においてトピック k が割り当てられる確率で，$\theta_{dk} \geq 0, \sum_K^{k=1} \theta dk = 1$ を満たす.

HDP HDP[10] は，LDA をノンパラメトリック化し，Dirichlet 過程によってトピック数の推定している．文書ごとの Durichlet 過程

$$G_d \sim DP(\alpha, H), d = 1, ..., D$$

は共通の基底分布 H を持ち，その基底分布は Dirichlet 過程

$$H \sim DP(\alpha_0, H_0)$$

から生成されたものと仮定する．また，LDA と同様の特徴ベクトル F_p で表す事ができる.

tfidf 法とトピックモデルの違いは，tfidf が対象文書に現れる語のみを特徴量とするに対して，トピックモデルは対象文書を確率的にトピックによって表すため，類似の語が対象文書に含まれていなくとも特徴として用いる事が可能である.

LDA と HDP の違いは，LDA では事前にトピック数を指定する必要があるが，HDP では最適なトピックを推定する事が可能である.

3.3 クラスタリング

特徴抽出によって得られた各パラグラフの特徴ベクトルを以下の手法を用いてクラスタリングする．python の機械学習ライブラリである scikit-learn を用いて各手法を実装した.

k-means 法 分割最適化クラスタリングの 1 つで，シンプルなクラスタリング手法で有名であるが，初期値依存性がある．以下の 4 ステップで k 個のクラスタに分ける事ができる.

(1) データ x_i，$(i = 1, ..., n)$ のクラスタをランダムに決める.

(2) 各クラスタの中心 $V_j(i = 1..k)$ を求める.

(3) 各 x_i と V_j の距離を以下の式によって求め，x_i のクラスタを最も近い中心のクラスタに更新する.

(4) （2），（3）を繰り返し，クラスタが更新されなくなるか，変化量が指定した閾値を下回った時に終了する.

Ward 法 凝縮型階層的クラスタリングの 1 つであり，はずれ値に強く，実用的だと考えられている．階層的クラスタリングとは近いクラスタ同士を順に併合していく手法であり，Ward 法ではクラスタ内での分散の変化が小さいクラスタを選択する．以下の 3 ステップでクラスタリングを行なう.

(1) すべてのデータに対して，別々のクラスタを割り当てる.

(2) クラスタ間の距離関数 $d(C_i, C_j), (i, j = 1...n, i \neq j)$ が最も小さいクラスタ同士を逐次的に併合する.

(3) すべてのデータが 1 つのクラスタに併合されるまで繰り返すと，デンドログラムという非終端ノードで表した二分木によって表示される階層構造を得る事ができる.

Ward 法で用いる距離関数は以下の式で表される.

$$d(C_i, C_j) = Var(C_{\cup} C_j) - (Var(C_i) + Var(C_j))$$

$Var(C)$：クラスタ C 内のデータの分散

デンドログラムの高さ（クラス間距離）によって切ることで，任意の数のクラスタに分ける事ができる.

k-means 法と Ward 法の違いは，分割するクラスタ数である．k-means 法はあらかじめクラスタ数を指定する必要があるが，Ward 法では，デンドログラム作成後にクラス間距離を閾値としてクラスタ数を指定せずにクラスタリングをする事が可能である.

4. 評価実験

比較表作成の最初の試みとして，単純でかつ類似したトピック構造である文書に対してクラスタリングを行い，その有用性を評価した.

Original Text: In China, the existence of a plethora of overlapping data privacy laws has traditionally made compliance very difficult for companies that collect personal information.

↓

[‘existence’, ‘plethora’, ‘overlapping’, ‘data’, ‘privacy’, ‘law’, ‘traditionally’, ‘made’, ‘compliance’, ‘difficult’, ‘company’, ‘collect’, ‘personal’, ‘information’]

図 3 前処理

Topic1	Topic2	Topic3
cost	provider	increasingly
operation	grow	economic
produce	global	efficiently
sector	reasonable	cross-border
rely	strategy	operation

表 1 トピック例

4.1 実験に用いるデータ

本実験では，European Centre for International Political Economy（ECIPE）の発表したデータの現地化要求（Data Localisation）についてのレポートを対象とする．国境を越えてデータを取り扱う事が増えたことで，自国のデータを海外で流通させる事を制限する Data Localisation が主要な課題の一つとして議論されている．本レポートは，Data Localisation がベトナム，韓国，インドネシア，中国に対して与える影響をまとめたものである．図のような pdf で提供され，文書は 8 から 9 パラグラフで構成されており，インターネットに関するトレンド（T），国内におけるデータ規制（D），経済への影響（I）の 3 つのトピックに分かれている．そこで，すべてのレポートをパラグラフごとに読み込み，3 つのクラスタに分割する．

クラスタリングを行なったラベルのすべての組み合わせを，人の手で作成した正解データと比較し，最も正解率の高かったラベルの正解率によって評価する．また，それぞれのレポートを 3 つのトピックに一致するように 3 パラグラフにまとめた単純なテストデータ（ECIPE simple）を作成し，オリジナルのテストデータ（ECIPE normal）と比較した．

4.2 実験結果

テストデータに対して前処理を行なった例を図 3 に示す．”laws”が”law”に統一されて，”In”などの一般語，”China”などの国に関する情報が削除されている事が確認できる．

LDA を 50 トピックで学習した結果の代表的トピックを図 1 に示す．また，HDP によって抽出されたトピック数の平均は 142 個となった．

各試行によって正誤率にばらつきがあるため，50 回試行した平均値の推移を図 4.2，図 4.2 に示す．

各特徴抽出手法，クラスタリング手法のすべての組み合わせに対して得られたスコアの平均値を表 2 に示す．

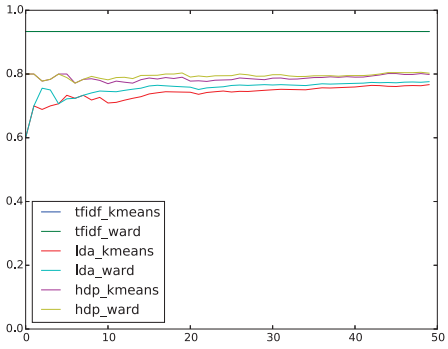

図 4 ECIPE simple

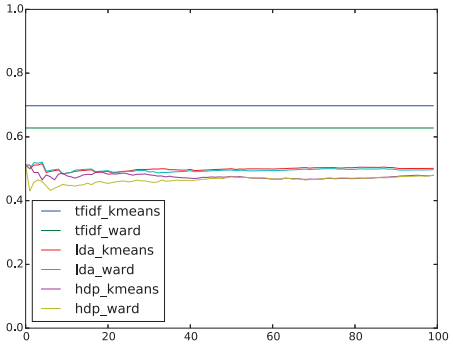

図 5 ECIPE normal

	simple		normal	
method	k-means 法	Ward 法	k-means 法	Ward 法
tf-idf 法	0.9333	0.9333	0.6976	0.6279
LDA	0.7766	0.8093	0.5011	0.4969
HDP	0.784	0.7926	0.4793	0.4767

表 2 スコア平均値

4.3 考察

ECIPE normal のように，同一文書内で一つのトピックが複数のパラグラフに分かれている場合のスコアが落ちる事が分かった．そこで，TexTTiling などのテキストセグメンテーション手法を用いて，同じ文書内での連続するパラグラフを同一トピックにまとめることで，精度を高める事ができると考えられる．また，今回実験で用いたレポートは一人の著者が複数のレポートについて書いたため，同一の構造になっており人にとってはトピックが理解しやすかった，図 6 のように同一の表現が多く，文書間の類似度が高かった．その為，LDA や HDP は潜在的なトピックを抽出する複雑なモデルであるため，シンプルな tfidf よりもスコアが低かったと考えられる．

本研究で目標とする文書は複数の著者によって記述された文書であるため，複数の著者によって記述された文書についても各手法の違いを評価する必要がある．幅広い種類の文書に対応するために，本稿ではラベルを決めずにパラグラフ間でのクラスタリングを行なったが，岡本ら [4] や

India: These findings show that the negative impact of disrupting cross-border data flows should not be ignored. The globalised economy has made unilateral trade restrictions a counterproductive strategy that puts India at a relative loss to others in the region, with no possibilities to mitigate the negative impact in the long run. If India fully enforces data localisation in all sectors, it will strongly impact the Indian economy by decreasing productivity, hampering exports and discouraging investment.

Vientam: These findings show that the negative impact of disrupting cross-border data flows should not be ignored. The globalised economy has made unilateral trade restrictions a counterproductive strategy that puts Vietnam at a relative loss to others in the region, with no possibilities to mitigate the negative impact in the long run. In addition to restricting online freedom, Decree 72 will heavily impact the Vietnamese economy by decreasing productivity, hampering exports and discouraging investment.

図 6 インド，ベトナム間の類似パラグラフ

Zirn ら [7] のように，国際公共政策に関するラベルの候補をあらかじめ作成しラベル付き文書で学習を行ない，分類を行なうことでスコアを向上させることが考えられる．

5. まとめ

本研究は文書中のトピックを比較項目とし，各国の文書に関して同一トピックでまとめた比較表の作成を目指す．本稿では，比較表作成の最初の試みとして，ECIPE の Data Localisation に関するレポートに対して教師無しで特徴を抽出し，クラスタリングを行い各手法を評価した．特徴抽出手法として，tf-idf 法，LDA，HDP，クラスタリング手法として，k-means 法，Ward 法を実装し，人の手によって作成した正解データと比較し，正解率を求めた．その結果，今回用いたレポートに対しては，最もシンプルな手法である tf-idf 法のスコアが最も高くなる事が分かった．これは，今回用いたレポートが同一著者によって作成されたものであったために，文書中の単語の種類が似すぎていたため LDA，HDP のスコアが tf-idf 法と比べて低くなったと考えられる．

今後の研究課題として，同一トピックに関する前後のパラグラフの統一，異なる著者によって作成された文書に対して実験を行なう．国際公共政策に関するラベル候補を作成し，ラベル付きの学習を行なった後に文書の分類を行なうアプローチも考慮すべきである．クラスタリングの次の課題は，同一トピックに関するパラグラフ間を比較し，内容の差異を抽出するである．同一トピックにおける各国の差異を可視化することで，より視認性を高める事ができると考えられる．多くの文書比較の研究が文書単位の関係の抽出であることに対し，本研究は複数文書内に存在する各トピック間の関係の抽出を目指す．

参考文献

[1] Daniel Bär1 Torsten Zesch and Iryna Gurevych. Text reuse detection using a composition of text similarity measures. In *Proceedings of COLING*, Vol. 1, pp. 167–184, 2012.

[2] Asli Celikyilmaz, Dilek Hakkani-Tur, and Gokhan Tur. Lda based similarity modeling for question answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pp. 1–9. Association for Computational Linguistics, 2010.

[3] 目黒光司, 笹野遼平, 榊原隆文, 菊池悠太, 高村大也, 奥村学. F ターム概念ベクトルを用いた特許検索システムの改良. 自然言語処理, 2015.

[4] 岡本和也, 竹村匡正, 黒田知宏, 長瀬啓介, 吉原博幸. 文脈に基づく類似診療文書検索システム. 生体医工学, Vol. 44, No. 1, pp. 199–206, 2006.

[5] Jonathan B Slapin and Sven-Oliver Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, Vol. 52, No. 3, pp. 705–722, 2008.

[6] Matthew J Gabel and John D Huber. Putting parties in their place: Inferring party left-right ideological positions from party manifestos data. *American Journal of Political Science*, pp. 94–103, 2000.

[7] Cäcilia Zirn. Analyzing positions and topics in political discussions of the german bundestag. In *Proceedings of the ACL 2014 Student Research Workshop, Baltimore*, pp. 26–33, 2014.

[8] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 248–256. Association for Computational Linguistics, 2009.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.

[10] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, Vol. 101, No. 476, 2006.