2K5-OS-25b-2

大規模パブリックコメントのトピック抽出に関する検討 「エネルギー・環境に関する選択肢に対する御意見の募集」を事例として

A study on the topics Extraction from Large-Scale Public Comments －As a case study of "public comment for choices on energy and the environment"－

岩見 麻子[*1] Asako IWAMI　木村 道徳[*2] Michinori KIMURA　松井 孝典[*3] Takanori MATSUI　熊澤 輝一[*4] Terukazu KUMAZAWA

[*1] 愛知工業大学 地域防災研究センター Aichi Institute of Technology Disaster Prevention Research Center

[*2] 滋賀県琵琶湖環境科学研究センター Lake Biwa Environmental Research Institute

[*3] 大阪大学大学院工学研究科 Graduate School of Engineering, Osaka University

[*4] 人間文化研究機構 総合地球環境学研究所 Research Institute for Humanity and Nature

In this study, the authors tried to examine the utility of the method called LDA (Latent Dirichlet Allocation) to extract the topics from large-scale public comments. As the result of case study with submitted comments of the "public comment for choices on energy and the environment" implemented by National Policy Unit in 2012, it was possible to clarify the problems such as necessity of developing an objective word selection method for identifying the issues being mentioned less frequently, and considering method for comparing the result of analysis with the issues defined by academic expert.

1. はじめに

環境政策の分野では住民参加が不可欠であり，広く採り入れられている手法としてパブリックコメント（以下，PC）がある．しかし PC は，提出された意見がどのように政策に反映されたかが不透明である点や意見に基づく議論の展開など，発展的手法が見られず形骸化している点など課題も指摘されている[山田 2011]．また，近年では数万件を超える大規模なものも現れている．たとえば，内閣官房国家戦略室が実施した「エネルギー・環境に関する選択肢に対する御意見の募集」には，88,634 件の意見が提出された．これら膨大な意見の全文は国家戦略室のウェブサイト上において PDF 形式で公開されている[内閣府 2012]．また，各意見をその内容によって事前に有識者から提起された 4 つの視点と細かな 25 の論点（**表 1** 参照）に分け，論点ごとに意見数を集計した結果も公表されていた．ただし，各意見を論点に分類した基準や具体的な集計方法などは公表されていない．このように，PC に提出された意見を客観的に把握あるいは集約し，活用するかは解決すべき課題である．

一方，大量のテキストからその文書の話題を表すトピックを把握する手法として LDA（Latent Dirichlet Allocation）[Blei 2003] が注目を集めている．そこで本研究では，前述した大規模 PC に提出された意見集合を対象に，専門家が提起した 25 の論点を，LDA によって機械的に再現できるかを検証することを目的とする．ここで，LDA で生成されるトピック集合の解釈について，各トピックを説明する単語数が多くなりすぎるとトピックの意味の可読性が低下して論点集合との対応の評価が困難になるため，本稿ではまず，LDA に入力する単語集合 W を最小化した上で、各パブリックコメントがより良く単一トピックに分類されるための方法を考察する．

2. 分析の枠組み

LDA を用いる場合，抽出するトピック数を分析者が任意に設定する必要があり，その決定方法について検討した研究も見られる[藤野 2014]．対象語数についても，語数が多すぎる場合，結果が複雑になりトピックを解釈することが困難であったり，比重が発散したりするのに対して，少なすぎる場合，文書全体のトピックを抽出するには不十分であることが考えられるため，これらのバランスを考慮した対象語数を慎重に検討する必要がある．

本稿では，まず 50～1,000 語まで，出現頻度順に $N_i = 50*i$ 個の単語を選定した出現単語の部分単語集合 W_i（i = 1~20）を作成し，W_i でコメント集合 C を Bag of Words 表現にして LDA を適用した．このとき，トピック数は国家戦略室が設定した論点の数と同じ 25 とした．次に，各 W_i を用いて生成したトピックモデル i について，各コメント c_j (j = 1~88,634) がトピック k（k = 1~25）に所属する事後確率分布を求め，その最大値 $\max(p_k|c_j)$ を全てのコメントに対して積算した $\Sigma_j \max(p_k|c_j) / N_j = \sigma_i$ を算出し，$\arg\max \sigma_i (W_i)$ を最適な部分単語集合 W_i と決定した．なお，σ_i は各コメント c_j を 25 のトピックのいずれかに分類する能

表 1 国家戦略室が設定した 4 つの視点と 25 の論点

視点	論点
原子力の安全確保と将来リスクの低減	① 原子力安全に不安，事故原因・影響も不明，健康被害もある
	② 核廃棄物は将来世代に負担を残す
	③ 原子力開発は倫理的に適切ではない
	④ 今脱原発か推進かを決められない，決めるべきでない
	⑤ 安全対策を強化することで，リスクを最少化できる
	⑥ 時間とコストがかかる廃炉を着実に進めることが重要
	⑦ 安全を担う人材と技術が必要である
	⑧ 原子力平和利用国としての責務を果たすべき
	⑨ 国家安全保障のため核関連技術を保有すべき
	⑩ 原発の不良債権化や立地地域への影響を懸念
エネルギー安全保障の強化	⑪ 再生可能エネルギーや新エネ開発こそ急ぐべき
	⑫ 国際エネルギー情勢を注視しいずれにも偏らず多様化を進めるべき
	⑬ 化石燃料の，調達源の多様化，戦略的活用が重要
	⑭ 非化石電源である原子力発電が重要
	⑮ 電力の安定供給のためには原子力発電が必要
	⑯ 今でも電気は足りている
地球温暖化問題解決への貢献	⑰ 温暖化対策にもっと積極的に取り組むべき
	⑱ 温暖化対策は他国の動向を見極めつつ推進すべき
	⑲ 温暖化対策は国外での実施に貢献すべき
	⑳ 温暖化対策は重視する必要ない
	㉑ 温暖化はしていない
コストの抑制空洞化防止	㉒ 新産業や雇用創出の好機である
	㉓ 経済への影響を見極めながら，エネルギーシフトすべき
	㉔ コストがあがり，経済に影響が出て，雇用が失われる
	㉕ エネルギー多消費産業の構造転換が必要となる

連絡先：岩見麻子，愛知工業大学地域防災研究センター，愛知県豊田市八草町八千草 1247，iwami-a@aitech.ac.jp

力が高いことを近似する指標である. その後, 部分単語集合 $\arg\max \sigma_i(W_i)$を構成する単語集合 w_iは, 事後確率の最大値 $\max(p_k|w_i)$により所属するトピック k を決定した. その上で, σ_iの値から W_i を選び, 各語について, それぞれのトピックに所属する事後確率分布を求め, その最大値 $\max(p_k|w_j)$のトピックに所属するものとして対象語を分類してみる.

なお, 形態素解析には ttm を, LDA には R の topicmodels パッケージを用いる.

3. 結果と考察

本稿では ttm の品詞体系の名詞のうち, 一般と固有名詞, サ変接続, 形容動詞語幹, ナイ形容詞語幹, 副詞可能, 複合名詞に分類された語を用い, 代名詞と英数字のみの 1 文字語は分析から除外した. なお, 対象品詞について出現語と出現回数を把握した結果, 異なり語数は 136,238 語, のべ出現回数は 2,654,039 回であった.

まず, W_iに対して LDA をそれぞれ適用し, $\arg\max \sigma_i(W_i)$を把握した結果を**図 1** に示す. 図には W_i ののべ出現回数が全出現回数に占める割合を併せて示している. 図に示すように, $\arg\max \sigma_i(W_i)$は, 語数が増加するにつれてその値が高くなる傾向が見られ, 特に 450 語と 800 語の時点において近似平均値が上昇していた.

次に, 各コメント c_j がトピック k に所属する事後確率分布の最大値の積算である σ_i に基づき語を分類した. ここでは, **図 1** に示した結果から, 450 語を用いて分類を試みた. その結果を**表 2** に示す. なお, 同 450 語(全出現語の約 0.33%), のべ出現回数は 1,334,136 回(同約 54.6%)であった. 表に示すように, 各トピックに分類された語数は, 2～42 語(最少:トピック 3 と 21, 最多:トピック 4 と 23)であった. 各トピックに含まれる語を見ていくと, 論点を表していると考えられるトピックも見られるが, 雑多な語を含み, あるいは語数が少なく論点の判断が困難なトピックも多く見られる.

この原因として, 対象語の選定に関する課題が考えられる. 本稿では名詞のうち代名詞と英数字のみの 1 文字語を対象から除外したのみで, 単純な出現頻度を用いて対象語を決定した. 対象語数と併せて, TFIDF など, 分析に意味のある語を客観的に選定する手法も用いて検討する必要がある. また, 対象語数についても, 多くなれば各語がそれぞれのトピックを意味する比重は減少し, トピック内の論点を推定することが困難になるため, 慎重に検討する必要がある.

4. おわりに

本稿では, 大規模 PC に提出された意見集合を対象に, 専門家が提起した 25 の論点を, LDA によって機械的に再現できるか検証を試みた. その結果, 次のような課題が明らかになった. まず, トピックを的確に推定するためには対象語数を慎重に検討する必要があることが確認できた. 本稿では単純な出現回数の上位語を用いて LDA を実施したが, 品詞を含め対象とする語の選定方法も併せて検討する必要がある. また, 本稿で対象とした大規模 PC は, 国家戦略室によって 25 の論点が設定されており, それとの比較によって分析結果の妥当性を検討することが可能であるが, その対応関係を比較する方法を検討することも課題として残されている. 加えて今後は, 語の分類結果を比較しより的確に論点を表すトピックを評価する方法や, 各意見の比重を用いたトピックの推定についても検討する必要があると考えられる.

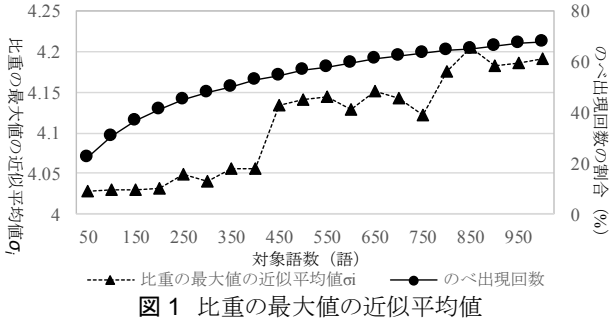

図 1 比重の最大値の近似平均値

表 2 語の分類結果(450 語)

1	選択, すべて, 現実, 化石.燃料, 活用
2	生活, 開発, 場合, 維持, 安全.神話, 普及, 利権, 電気.代, 不足, 報道
3	管理, 本当
4	選択肢, 反対, 絶対, 福島.原発.事故, 議論, 福島.原発, 負担, 声, 研究, 原因, 世界中, 情報, 故郷, いま, 月, 無視, 教訓, 無責任, 検討, 考え, 災害, 話, 様々, 拡大, 使用.済み.燃料, 燃料, 被爆, 保障, 確実, 方針, 自体, 電力.供給, 現時点, 放出, 構築, 専門.家, 最後, 存続, 見通し, 確認, 風力.発電, 長期.的
5	原子力, 核, 自分, 目, 火力.発電, コントロール, 経験, 明確, 心配, 言葉, 自分.たち, 将来.的, 比率
6	政府, 廃.炉, 現在, 推進, 人類, 社会, ゴミ, お金, 不可能, 方々, 一部, 利益, 判断, 火力, 豊か, 証明, 疑問, システム, 国家, 現実.的, 知恵, エネルギー.源, 住民, 健康.被害, 家庭
7	国, 経済, 状況, 地球, 節電, 企業, 状態, 十分, 人.たち, 生命, 制御, 雇用, 電力.不足, 海外, 時代, 広島, 長崎, 被曝, 信頼, 原発.推進, 発電.所, 段階
8	再生.可能.エネルギー, 私.たち, 人々, 意見, 依存, 世代, 支持, 前, 優先, 原子.炉, 積極.的, 具体.的, 天然.ガス, 崩壊, 原子力.エネルギー, 提案, 生産, 発送.電.分離, 石炭, 耳, 真摯, 再生, 加速, 誇り, 温室.効果.ガス, 要求, 核.エネルギー, エネルギー.効率, きれい, 約束, エネルギー.計画, 気候.変動, 削減.目標, 国際.公約, エネルギー.部門
9	福島, 国民, 問題, 人間, 責任, 転換, 発生, 政治.家, 東電, 信用, 夏, 唯一, 行動, 立場, 中国, 明白
10	稼働, 多く, 放射.性.物質, 先, 太陽光, シフト, 事態, 無理, 大変, 規模, クリーン, 原発.依存.度, 破綻, 施設
11	日本, 膨大, 実施, 形, 外国, 事故.後, 想像
12	地震, 安全.性, 努力, 代替.エネルギー, 想定, 取り返し, 処分, 速やか, 地震国.日本, 即時
13	安全, 人, 解決, 子供, 思い, 賛成, チャンス, あと
14	日本人, 実現, 原発.依存, 間, 日本.国民, 歴史
15	放射能, 技術, 力, 不安, 利用, 核.廃棄.物, 大飯原発, 全て, 危険.性, 太陽光.発電, 東日本.大震災, 破壊, 一刻, 安定, 共存, 技術.開発, 時期, 核.燃料.サイクル, 移行, 安価, 実行, 幸せ, 協力, 面
16	自然.エネルギー, 処理, 今回, 現状, 使用.済み.核.燃料, 政策, 風力, 使用, 時間, 前提, 重要, 子供.達, 早急, 期待, 次, CO, 税金, 発電.方法, 大量, 莫大, 水力, 認識, 導入, 設置, 考慮, 家族, 他国, 観点, 官僚, 目標, 提示, 不便, 課題, 排出, 輸出, 省エネルギー, 増加, 国策, 悪影響, 結論
17	事故, 影響, 原子力.発電.所, 汚染, 国土, 子ども.たち, 処理.方法, 費用, 事実, 放射能.汚染, 国内, 決定, 昨年, 甚大, 私.達, 子ども, 一つ, 主張
18	シナリオ, 危険, 原子力.発電, 電力, リスク, 電力.会社, 稼動, 発電, 確立, 理解, 省エネ, 活.断層, 暮らし, 原爆, 我が国, 保証, 継続, 多大, 保管, アメリカ, 中心
19	今, 命, 大切, 土地, 遺産, 犠牲, 地震.大国, 大事, 反省, 運転, 便利, 孫, 姿勢, 被爆.国, 気持ち, 金
20	今後, 可能.性, 確保, 廃止, 対策, ドイツ, 電気.料金, 説明, 核.燃料, 避難, 同様, 経済.成長, 指摘, 一番
21	未来, 自然
22	方法, 放射.性.廃棄.物, 存在, 子供.たち, 海, 無駄, 過去, 目先, 真剣, 放射線, 効率, 仕方
23	世界, 原発.事故, 将来, 電気, 被害, 明らか, エネルギー.政策, コスト, 他, 可能, 対応, 手, 停止, 津波, 収束, 新た, 負, 発展, 経済.的, 地震国, 地熱, 想定.外, 技術.力, 決断, 仕事, 困難, 供給, 建設, 震災, 納得, 人災, 安全.対策, 自然.災害, 恐怖, 次世代, 深刻, 即刻, 間違い, 投資, 東京電力, クリーン.エネルギー, 持続.可能
24	原発, エネルギー, 理由, 安心, 環境, お願い, 希望, 健康, 意味, 道, 方向, 結果, 子孫, たくさん, 政治, 人達, 心, 後世, 覚悟, 根拠, 日本.経済
25	必要, 地域, 廃棄.物, 産業, 場所, 完全, 非常, 資源, 水, 予算, チェルノブイリ, 削減, もんじゅ, 石油, ウラン, 地熱.発電, 再生.エネルギー, 家, レベル, 地球.温暖.化, 早期, 地球.上, 視点, 機会, フクシマ

謝辞

本研究は公益財団法人 日本生命財団の環境問題研究助成を受けたものである.

参考文献

[山田 2011] 山田久美子, 柳下正治: 我が国の気候変動政策における意思決定プロセスへの市民関与の発展, 環境科学会誌, Vol.24, No.5, pp.422-439, 2011.

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

[内閣府 2012] 内閣官房国家戦略室 エネルギー・環境会議: ―政策―エネルギー・環境会議 パブリックコメント, <http://www.cas.go.jp/jp/seisaku/npu/policy09/archive11.html>, (2013.1.28 参照)

[藤野 2014] 藤野巌, 星野祐子: LDA 法におけるトピック数の決定法およびトピックの評価法について : Twitter ストリーミングデータを応用例として, 電子情報通信学会技術研究報告. DE, データ工学, Vol.114, No.101, pp.67-72, 2014.