確率的潜在意味スケーリング

持橋 大地[1,a)]

概要：テキストをある意味的な軸に沿って連続的に測ることは, 社会科学・人文科学などにも様々な応用を持つ重要な問題である. 本研究ではこのために, 項目反応理論とニューラル単語ベクトルに基づいて, 政治学方法論の分野で提案された潜在意味スケーリング (LSS) の考え方を統計モデルとして実現する, 確率的潜在意味スケーリング (PLSS) を提案する. また, 分析対象となるテキストをキーワードに基づいて選択するための枠組として, 潜在トピックモデルおよびニューラル文書ベクトルに基づく二種類の統計的な方法を提案する. 政治学分野での公開データを用いた実験により, PLSS は LSS より人間の評価と高い相関を見せることを確認し, また LSS を確率モデルとして捉えることで, 様々な統計的拡張を可能にする.

キーワード：項目反応理論, 政治学方法論, 潜在意味解析, 単語ベクトル, ガウス–エルミート求積

Probabilistic Latent Semantic Scaling

Daichi Mochihashi[1,a)]

Abstract: Measuring text in a continuous scale is a fundamental problem that has many applications including social sciences and humanities. This paper proposes a probabilistic extension to Latent Semantic Scaling (LSS) in political methodology, called Probabilistic Latent Semantic Scaling (PLSS). Leveraging Item Response Theory and neural word vectors, we show that PLSS consistently outperforms LSS in terms of relevance to human evaluation. We also propose two statistical methods to select texts to be analyzed that are associated with the given keywords. Probabilistic treatment of LSS enables many future extensions, including handling missing data and time series analysis.

Keywords: Item Response Theory, Political Methodology, Latent Semantic Analysis, Word vectors, Gauss-Hermite quadrature

1. はじめに

テキストをある尺度 (スケール) で連続的に測りたい, という場面に我々はよく遭遇する. たとえば, 自然言語処理や心理学において, あるレビューやアンケートがどれくらいポジティブなのか, どれくらいネガティブなのか知りたい場面はよく存在するし, 社会科学においてはある法案がどれくらい左翼=革新的か, あるいは右翼=保守的かを知ることで, この法案に賛成する議員の立場を逆算して焙り出すことが可能になる [1][2]. 特に, テキストの極性自体はお客様センターへの苦情や自民党から提出された法案といった形で自明な場合があり, そこではむしろ, どれくらい強い苦情なのか, どれくらい保守的な法案なのかといった「程度」を知ることが実用上重要となると考えられる.

しかし, こうした問題は, 1/0 の分類を行う従来のテキスト分類や極性判定のような教師あり学習では対応することができない. ロジスティック回帰のような分類器は, 尤度を上げるために結果を 1 か 0 のどちらかの極に寄せる傾向があり, SVM のようなベクトル空間で動く分類器においても, 分類平面からの距離がクラスに対応する尺度であることは保証されていないからである [3]. また, そもそもどういった尺度で違いがあるのかを知りたい探索的 (exploratory) な分析の場合は, 事前に人が固定された観点で付与した教師データは意味をなさない.

特に社会科学においては, 分類問題に還元できないこうした問題は重要であり, 中でも統計学や機械学習の導入が最近急速に進んでいる政治学 (政治学方法論; political methodology) の分野では, Wordfish [4] あるいは LSS [5][6] が, この目的でテキストを分析するために使われている. 後で説明するように, Wordfish は教師なしモデルであるが, LSS はキーワードの形で分析者が尺度の視点を決めることのできる方法であり, R パッケージ[*1]もあるために世界的に広く使われている.

しかし, LSS は古典的なベクトル空間モデルによる潜在意味解析に基づいているために, 推定値の分散が非常に大きく, また確率的な解釈や拡張が難しいという実用的な問題を抱えている. 実際に, 5 章で示すように, LSS の尺度との相関はかなり弱いものになっている. また, Wordfish と LSS のいずれにおいても, 分析の対象となるテキストを絞り込むための系統的な方法論がなく [4], 分析の際の障害となってきた.

そこで本研究では, テキストをこうした連続値で測るために, 心理統計学における項目反応理論 (IRT) をもとに,

[1] 統計数理研究所 数理・推論研究系
The Institute of Statistical Mathematics
[a)] daichi@ism.ac.jp

[*1] https://github.com/koheiw/LSX. CRAN への登録上, 名前が LSX になっていることに注意されたい.

Word2Vec 等で得られる現代的なニューラル単語ベクトルを利用し, LSS の考え方を統計モデルとして実現する確率的潜在意味スケーリング (PLSS) を提案する. 分析の前段においても, 背景分布を持つ LDA [7] およびニューラル文書ベクトル [8] を利用して, 各テキストが分析したいキーワード集合に関係する「関係確率」を求め, 上記のニューラルな IRT と組み合わせることで, PLSS が様々な話題を持つテキストの中からキーワードに関係する文書を自動的に抽出し, 人間の評価ともより高い相関を持つ尺度値を出力できることを示す.

本論文は, 以下のように構成される. まず 2 章で, テキストの尺度分析のための Wordfish と LSS について解説する. 3 章で本論文の基本となる項目反応理論 (IRT) と提案する PLSS のモデル化について説明し, 学習データの各テキストの潜在尺度 θ を数値的に積分消去するための適応的なガウス–エルミート求積に基づく学習法について述べる. また, 4 章で分析のためにコーパスからキーワードに関係するテキストを確率付きで抽出する方法について説明する. 5 章で政治学において尺度を判定した公開データにおける実験を行い, 提案手法の優位性を示す. さらに実際に中国共産党の潜在的なイデオロギーの時間変化と国会議事録の分析に適用した例を示し, 6 章でまとめと今後の課題について述べる. 提案する PLSS の R パッケージは, `github` において公開予定である.

2. テキストの尺度分析と統計モデル

政治学において法案や演説, 議事録といったテキストの分析は重要な課題であり [9], さまざまな方法が提案されてきた. テキストマイニングの分野で行われているように, 単純に単語ベクトルの和の第一主成分をプロットするといったヒューリスティックな方法もあるが, より系統的な方法として, Slapin らによる教師なし手法である Wordfish [4] および渡辺による半教師あり手法である LSS [5][6] が広く使われている.

2.1 Wordfish

Wordfish はテキストの系列[*2]から極性の変化を発見することのできる教師なし学習の統計モデルであり, ポアソン分布に基づいた生成モデルになっている. y_{ivt} を政党 i が時刻 t のテキスト (たとえば選挙時の公約) で単語 v を使った頻度とすると, 観測値 $Y=\{y_{ivt}\}_{i,v,t}$ について, Wordfish は次の確率を最大化するパラメータ $(\alpha,\beta,\theta,\phi)$ を求める.

$$\begin{cases} p(Y)=\prod_i\prod_t\prod_v \mathrm{Po}(y_{ivt}|\lambda_{ivt}) \\ \lambda_{ivt} \;=\exp\left(\alpha_{it}+\beta_v+\phi_v\cdot\theta_{it}\right) \end{cases} \quad (1)$$

ここで α_{it} はテキスト it での固定効果, β_v は単語 v の固定効果で, 興味があるのは単語 v の極性軸上での位置 ϕ_v および, 政党 i の時刻 t での潜在位置 (理想点) θ_{it} である.

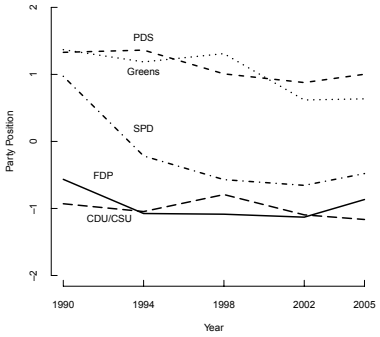


図 1 Wordfish によって推定された, ドイツの各政党の外交方針 θ_{it} の時間変化の例 (1990–2005) ([4] より引用).

式 (1) は, 頻度 y_{ivt} はポアソン分布に従うが, その期待値=分散 λ_{ivt} はテキスト it と単語 v で決まるベースライン $\alpha_{it}+\beta_v$ を, 政党の持つ極性 θ_{it} と単語の持つ極性 ϕ_v で上下して決まるということを表している. $\theta_{it}>0$ が右翼を表すとき, 政党の位置と単語の位置の符号が一致する, すなわち単語 v (たとえば “軍備”) の位置も同様に $\phi_v>0$ であるか, または $\theta_{it}<0$ すなわち左翼で $\phi_v<0$ のときに単語 v (たとえば “社会保障”) の期待値は増加する. 逆に θ_{it} と ϕ_v の極性が異なるときには, 単語 v の期待値は減少することになる. 推定には EM アルゴリズムを用い, 適切な初期値から始めて, α および θ の推定と β および ϕ の推定を反復する.

Wordfish は系統的な統計モデルであるが, この方法で発見される θ の極性軸が分析の目的と対応している保証はない. 適切な分析のためには, 入力テキスト集合を注意深く選択する必要があるが, その方法は Wordfish には含まれておらず, 客観的な方法は示されていない. また, Wordfish は「与えられたテキストの頻度」を説明するものであり, 学習される式 (1) の α_{it} は暗黙にテキストの長さに依存しているため, 学習したモデルを新しいテキストに適用できないという問題点もある.[*3] Wordfish によって推定したドイツの政党の θ_{it} の推定例 [4] を図 1 に示した.

2.2 Latent Semantic Scaling (LSS)

これに対して, 渡辺が政治学方法論の分野で開発した LSS (潜在意味スケーリング)[5][6] は, 「目標語 (target words)」という形で分析の対象を, 「種語 (seed words)」という形で尺度の視点をそれぞれ指定することのできる半教師ありアルゴリズムである.

目標語と特徴語 LSS ではまず, 指定した目標語 (target words) を使い, 分析に用いる語彙を特徴語 (entry words) として決定する. これには目標語 (たとえば “econom*” や “政治”) から一定の窓内で共起した語のうち, 符号付き χ^2 検定値の小さい順にたとえば 2,000 語を用い [6], “growth, social, recovery, $\cdots$” のような語彙を特徴語とする. また, 文書–単語の共起頻度行列を SVD で分解することで, 潜在意味解析 (LSA)[12] に基づいて各目標語および特徴語の単

[*2] これらの方法では時系列性を直接モデル化していないが, 本研究のように確率モデルとして捉えることで, ガウス過程 [10] などを用いた時系列拡張を行うことも可能になる.

[*3] ポアソン分布をテキストの長さ (一般に既知) で条件づければ多項分布が得られるため [11], 式 (11) のようにモデルを多項分布で再定式化すれば, 新しい文書にも適用することができる.

語ベクトルを計算しておく.

種語　分析の視点として, 極性 $p[s]$ (たとえば 1 や −1) の付いた単語リストを種語 (seed words) として与える. 次に, 上記の各特徴語 f について, その極性 $p[f]$ を種語 s とのコサイン距離により, 次のように計算する.

$$p[f] = \frac{1}{|S|} \sum_{s \in S} \cos(\overrightarrow{f}, \overrightarrow{s})\, p[s] \tag{2}$$

ここで S は種語の集合で, $\overrightarrow{f}, \overrightarrow{s}$ はそれぞれ f および s の単語ベクトルである. この様子を図 2 に示した.

テキストの極性の計算　$p[f]$ が式 (2) で計算できると, 与えられたテキストの極性は, テキストのうち特徴語だけを抜き出したものを d として

$$p[d] = \frac{1}{|d|} \sum_{f \in d} n(f)\, p[f] \tag{3}$$

として計算される. ここで $n(f)$ は特徴語 f の d 内での頻度, $|d|$ は d の長さである. この値は平均や分散が一定の値にならないため, 平均 0, 分散 1 となるように正規化した結果を LSS の最終的な出力とする.

LSS は, それまで使われていた人手による極性単語リスト (たとえば Lexicoder Sentiment Dictionary, LSD [13]) による分析を自動化し, 単語リストが存在しない言語にも適用できるようにしたもので, R パッケージにより政治学の多くの研究で使われている [6]. 図 3 に, [6] で New York Times の 10,000 記事を標準的な正負の種語を使って分析した例を示した.

LSS は有用なアルゴリズムであり, 各時期での文書の極性の平均値は人手による評価とよく一致することが示されている [6]. しかし一方で, 図 3 の各点にみられるように, 各文書の極性は非常にばらつきが大きいという問題が原論文によっても指摘されている. また, Word2Vec や GloVe のような現代的なニューラル単語ベクトルではなく, ベクトル空間モデルによる古典的な潜在意味解析を用いているために, ストップワードや頻度の前処理といったヒューリスティックおよびその際のパラメータに依存することに加え, 5 章に示すように, ベクトル空間モデルで得られる極性の「方向」が, 必ずしも人間の期待するものと一致しないという問題点がある.

そこで本研究では, こうした潜在的な「尺度 (スケール)」に対応する連続値を扱うことのできる方法論である項目反応理論 (IRT) をもとに, Wordfish のモデルを参考に LSS を

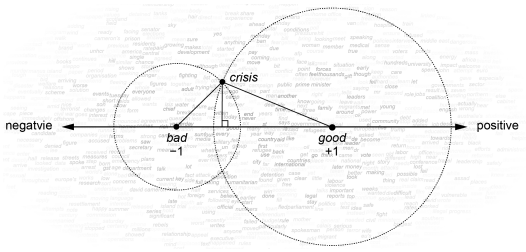


図 2　LSS による特徴語の極性の計算 [5]. 種語とのコサイン距離に基づき, 分析に用いる特徴語の極性をスコアリングする.

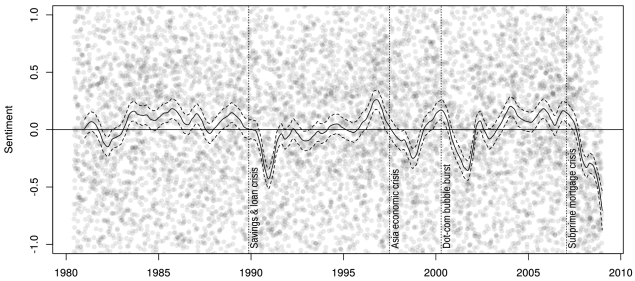


図 3　LSS によって推定した, New York Times の記事の極性の時系列変化 [6]. 点は各文書の極性を, 実線はスムージングしたそれらの平均を表している.

統計モデルとして確率化し, これらの問題点を克服することを試みる.

3. 確率的潜在意味スケーリング

項目反応理論 (item response theory, IRT) [14] は心理統計学の分野で開発された方法論であり, テストの結果から被験者の潜在的な能力あるいは傾向 θ を推定することを目的とする. 各被験者の θ が標準正規分布 $\mathcal{N}(0,1)$

$$\theta \sim \mathcal{N}(0,1) \tag{4}$$

に従っているとすると, 最も基本的な 2PL(2 パラメータロジスティック) モデルでは, 被験者 i の問題 j に対する正答確率 p_{ij} を次のようにモデル化する.

$$p_{ij} = \sigma(a_j(\theta_i - b_j)) \tag{5}$$

ここで $\sigma(x)=1/(1+e^{-x})$ はロジスティック関数であり, 式 (5) は, 問 j への正答率は問題の難易度を表す b_j を境に, a_j に比例する確率で上昇することを表している. 被験者 i の能力 (あるいは傾向) θ_i が b_j に比べて充分大きければ式 (5) の正答率は 1 に近づき, 小さければ 0 に近づく. 観測データとして $Y=\{y_{ij}\}_{i,j}$ を考え, y_{ij} は被験者 i が問題 j に正答したとき 1, 誤ったとき 0 とすると, Y の確率は

$$p(Y|a,b,\theta) = \prod_i \prod_j \mathrm{Bernoulli}\,(y_{ij}|p_{ij}) \tag{6}$$

$$= \prod_i \prod_j \mathrm{Bernoulli}\,(y_{ij}|\sigma(a_j(\theta_i - b_j))) \tag{7}$$

と表すことができる. IRT では, 事前確率 (4) の下で尤度 (7) を最大にするパラメータ $\{a_j, b_j\}$ を求め, 同時に被験者の能力 $\{\theta_i\}$ を推定する.

標準的な IRT はテストに対する正解/不正解の二値データ y_{ij} に対するベルヌーイ分布の統計モデルであるが, 式 (7) と式 (1) を比較するとわかるように, Wordfish はポアソン分布に基づく IRT の一種とみなすことができる. ただし, Wordfish では θ が標準正規分布に従うことは仮定されていないという違いがある.

3.1 確率的潜在意味スケーリング

そこで本研究ではまず, 式 (5) を多項分布に拡張する形で, あるテキストにおける単語 v の確率を次のようにモデル化する.

$$p(v|\theta,\phi) \propto p(v)\exp(\theta\cdot\phi_v) \tag{8}$$

$$= \frac{\exp(\log p(v)+\theta\cdot\phi_v)}{\sum_{v=1}^{V}\exp(\log p(v)+\theta\cdot\phi_v)} \tag{9}$$

ここで $\phi_v \in \mathbb{R}$ は式 (1) の Wordfish の場合と同様に, 単語 v の「極性」を表すパラメータである. θ はこのテキストのもつ潜在的な極性で, 標準正規分布 $\mathcal{N}(0,1)$ に従うとする.

式 (8) より, このモデルでは単語 v の確率は ϕ_v と θ の正負と大きさが一致すれば事前確率 $p(v)$ より高くなり, 逆になれば低くなる, というモデルになっている. $p(v)$ はコーパスから容易に計算できるため, このモデルは IRT の多項分布化, あるいは式 (9) で表される多値ロジスティック回帰において, θ も ϕ も未知の場合の教師なし学習とみなすことができる.

このとき, テキスト d の確率は単語 v のテキスト内での頻度を n_{dv} とおくと

$$p(d|\theta,\phi) = \prod_{v=1}^{V} p(v|\theta,\phi)^{n_{dv}} \tag{10}$$

と書けるから, D 個のテキストからなるコーパス $\mathcal{D}$ 全体の確率は

$$p(\mathcal{D}|\Theta,\phi) = \prod_{d=1}^{D}\prod_{v=1}^{V} p(v|\theta_d,\phi)^{n_{dv}} \tag{11}$$

と表される. Wordfish と同様に, 事前確率 (4) の下で式 (11) を最大化するパラメータ $\Theta=\{\theta_1,\cdots,\theta_D\}$ および $\phi_1,\cdots,\phi_V$ を最適化や MCMC 法などにより計算することができる.[*4]

ただし, こうするとたとえ単語 v と w が意味的に関係が深くても, ϕ_v と ϕ_w は別のパラメータとして推定しなければならないという問題がある. たとえば $\phi_{\text{good}} > 0$ と推定できても, これは $\phi_{\text{excellent}}$ や ϕ_{well} とは無関係であり, excellent や well がコーパスに現れなければ, まったく学習することができない.

そこで, $\phi_1,\cdots,\phi_V$ を独立に学習する代わりに, 与えられたコーパスあるいは一般的なコーパスから事前に Word2Vec や GloVe 等で学習された K 次元のニューラル単語ベクトル $\overrightarrow{v}$ を用いて, ϕ_v を

$$\phi_v = \boldsymbol{\beta}^T\overrightarrow{v} \tag{12}$$

とモデル化する. これにより, V 個の独立な $\phi_1,\cdots,\phi_V$ を求める代わりに, K 次元の係数ベクトル $\boldsymbol{\beta}$ を一つだけ推定すればよい. このとき, 式 (9) は $\ell_v=\log p(v)$ とおけば

$$p(v|\theta,\boldsymbol{\beta}) = \frac{\exp(\ell_v+\theta\cdot\boldsymbol{\beta}^T\overrightarrow{v})}{\sum_{v=1}^{V}\exp(\ell_v+\theta\cdot\boldsymbol{\beta}^T\overrightarrow{v})} \tag{13}$$

と表すことができ, 式 (2) の LSS のコサイン類似度 ($\simeq$ 内積) の確率モデル化になっているといえる.

この $\boldsymbol{\beta}$ は, 単語埋め込みベクトルの空間において “良い–悪い”, “右翼–左翼” といった θ の極性を与える「極性軸」, あるいは「意味方向」を表している. $\boldsymbol{\beta}$ は完全に教師なしで学習することもできるが, 2.1 節で述べたように, そうして得られた $\boldsymbol{\beta}$ が分析の目的と一致しているとは限らない. そこで, 本研究では表 1 のように, 正例および負例として与える少数の極性語辞書から, 単語ベクトルを用いて $\boldsymbol{\beta}$ を次のように計算する. なお, $\boldsymbol{\beta}$ のノルムは 1 に正規化する.

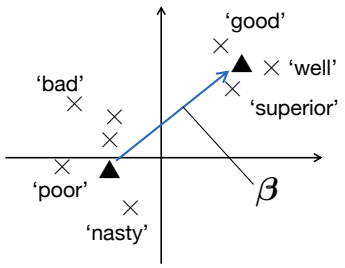


図 4 極性辞書と単語ベクトルによる $\boldsymbol{\beta}$ の計算.

$$\boldsymbol{\beta} \propto \left(\frac{1}{|S_+|}\sum_{v\in S_+}\overrightarrow{v} - \frac{1}{|S_-|}\sum_{w\in S_-}\overrightarrow{w}\right) \tag{14}$$

ここで S_+ は極性辞書のうち正の語の集合, S_- は負の語の集合とした. 式 (14) は単純に, $\boldsymbol{\beta}$ として負の単語ベクトルの平均から, 正の単語ベクトルの平均へ向かう方向を取ることを表している. この様子を図 4 に示した. 極性辞書としては, 表 1 に示した LSS に付属する標準極性辞書 [15] のように, ごく少数の単語のリストがあればよい. なお, LSS と異なり, 正例・負例の単語が「最も極端な語」である必要はなく, 意味方向のみが合っていればよいことに注意されたい.

ここで「正例」「負例」とは必ずしも感情的な正負と関係していなくてもよく, 「右翼–左翼」「都会–田舎」「東洋–西洋」のように, 任意の意味的な軸を扱うことができる. Schütze らは, 単語埋め込みの空間に実際にこうした, 与えられたタスクに関係する低次元の部分空間が存在することを発見し, これを超密埋め込み (Ultradense embedding) と呼んでいる [16]. 超密埋め込みは, 最も単純には, この場合のように 1 次元の部分空間となる. ただし, 予備実験でこの超密埋め込みを行列の固有ベクトルとして求める DensRay [17] を使用したところ, 式 (14) と比べて明らかにノイズの多い方向となったため, 本研究では単純な式 (14) を採用することとした.[*5]

図 5 に, 表 1 の標準的な極性辞書と, 160 万文の New York

表 1 LSS に付属する標準的な極性辞書 [15]. + が正例を, − が負例を表す.

+	good, nice, excellent, positive, fortunate, correct, superior
−	bad, nasty, poor, negative, unfortunate, wrong, inferior

[*4] これは多項モデルであるため, 式 (1) の Wordfish と異なりパラメータが学習テキストの長さに依存せず, 新しいテキストについても適用することができるという特徴がある.

[*5] これは, DenseRay の目的関数が, 極性辞書で単語 v の属する符号を $s(v)$ としたとき, 行列 $\mathbf{A}=\frac{1}{N}\sum_{\substack{(v,w):\\ s(v)=s(w)}}(\overrightarrow{v}-\overrightarrow{w})(\overrightarrow{v}-\overrightarrow{w})^T - \frac{1}{M}\sum_{\substack{(v,w):\\ s(v)\neq s(w)}}(\overrightarrow{v}-\overrightarrow{w})(\overrightarrow{v}-\overrightarrow{w})^T$ の固有ベクトルの計算に帰着され, 正例と負例の内部およびその間のペアを結ぶ個別のベクトルの和になっていることが原因だと考えられる. 式 (14) と異なり, この定式化では S_+, S_- の中で和をとることでノイズを消す作用が働かず, v,w の選択に起因するノイズが $\mathbf{A}$ に残ってしまうからである.

nice	0.5039	tasting	0.4080	accommod..	0.3890
versatile	0.5016	sturdy	0.4073	comfortable	0.3875
honored	0.4739	chosen	0.4062	amenities	0.3868
excellent	0.4648	invited	0.4040	flexible	0.3857
fortunate	0.4606	luxurious	0.4021	craft	0.3824
terrific	0.4522	exciting	0.3996	grateful	0.3799
pleasant	0.4485	thankful	0.3992	thrilled	0.3794
wonderful	0.4396	suite	0.3966	agassi	0.3767
selected	0.4391	great	0.3932	prepared	0.3756
happy	0.4295	unique	0.3917	ready	0.3745
coveted	0.4156	available	0.3914	able	0.3732
perfect	0.4125	fine	0.3910	cellars	0.3706

(a) 極性 $\phi_v > 0$ の上位単語.

famine	−0.5043	riots	−0.4519	ignoring	−0.4317
rioting	−0.4987	racism	−0.4510	beatings	−0.4309
mismana..	−0.4945	fear	−0.4499	rebellion	−0.4296
rampant	−0.4871	unneces..	−0.4482	violence	−0.4294
greed	−0.4800	panic	−0.4466	provoking	−0.4293
crippling	−0.4735	strife	−0.4465	abruptly	−0.4277
mongering	−0.4710	chaos	−0.4444	subprime	−0.4265
unrest	−0.4657	abuses	−0.4364	fears	−0.4252
blaming	−0.4609	carnage	−0.4357	violently	−0.4246
hysteria	−0.4604	tolerating	−0.4355	violent	−0.4179
unchecked	−0.4582	scandals	−0.4334	disastrously	−0.4175
intimida..	−0.4521	collapse	−0.4332	plague	−0.4171

(b) 極性 $\phi_v < 0$ の下位単語.

図 5 New York Times のコーパスから事前学習した Word2Vec と表 1 の極性辞書を用いて計算した単語の極性 ϕ_v.

Times の記事から Word2Vec で学習した $K = 100$ 次元の単語ベクトルを用いて計算した単語の極性 $\phi_v = \boldsymbol{\beta}^T \overrightarrow{v}$ を示した. 非常に少ない教師データにもかかわらず, 肯定的な単語および否定的な単語が, その強さとともに連続的に取り出せている様子がわかる.

このモデルは LSS の確率化とみなせるため, **確率的潜在意味スケーリング** (Probabilistic Latent Semantic Scaling, PLSS) と呼ぶ. 極性辞書を用いる場合は, PLSS は単語ベクトルの計算以外にパラメータの学習を必要としない. 式 (10) から, テキスト d と極性 θ の同時確率は

$$p(d, \theta) = \prod_{v=1}^{V} p(v|\theta, \boldsymbol{\beta})^{n_{dv}} \cdot p(\theta) \tag{15}$$

$$= \prod_{v=1}^{V} \left(\frac{\exp\left(\ell_v + \theta \cdot \boldsymbol{\beta}^T \overrightarrow{v}\right)}{\sum_{v=1}^{V} \exp\left(\ell_v + \theta \cdot \boldsymbol{\beta}^T \overrightarrow{v}\right)} \right)^{n_{dv}} \cdot \mathcal{N}(\theta|0, 1) \tag{16}$$

であり, これを最大にするテキストの潜在的な極性 θ の MAP 解は, 1 次元の最適化で容易に計算することができる.

3.2 PLSS の半教師あり学習

上では θ の極性を表す基準として LSS と同様に少量の極性辞書を用いたが, こうした辞書が分析対象について自明に作成できるとは限らない. 例えば, 欧州において移民労働者についての賛成派と反対派を特徴づけるキーワードが, 分析前から明らかとは限らないからである.

しかし, そうした場合でも典型的な「正例」のテキストと「負例」のテキストは示せる場合が多いと考えられる. 直感的には, それぞれのテキストに共通して現れる単語 (単語ベクトル) から, 間接的に単語の極性が導かれるはずである. コーパスのうち, こうした極性が既知のテキストの集合を X_ℓ, それらへの 1/0 のラベルを Y_ℓ とし, それ以外の極性が未知のテキストの集合を X_u とおくと, 最も簡単には, Y_ℓ, X_ℓ, X_u の同時確率を

$$p(Y_\ell, X_\ell, X_u) = p(Y_\ell|X_\ell)p(X_\ell)p(X_u) \tag{17}$$

と定義し, この確率を最大化する PLSS のパラメータ $\boldsymbol{\beta}$ を学習すればよいと考えられる.

しかし, 予備実験によりこの方法は上手く行かないことがわかった. 式 (17) の対数をとると

$$\log p(Y_\ell, X_\ell, X_u) = \log p(Y_\ell|X_\ell) + \log p(X_\ell) + \log p(X_u) \tag{18}$$

となるが, この第 1 項・2 項の教師ありデータのテキストの対数尤度より, 第 3 項の教師なしデータの対数尤度の方が圧倒的に大きいため, 単純に式 (18) を最大化すると教師ありデータがほとんど無視されて, 教師なしデータの尤度のみを最大化する $\boldsymbol{\beta}$ が学習されてしまうからである. この結果, 極性の尺度のために与えた教師ありデータはほぼ意味を持たず, Wordfish のように教師なしでコーパスを説明する極性軸 $\boldsymbol{\beta}$ が学習されてしまう.

そこで本研究では, 式 (18) の第 1・第 2 項の (Y_ℓ, X_ℓ) だけを用いて $\boldsymbol{\beta}$ を学習することとした.[*6] すなわち, 教師データの各テキストとそのラベル $(y, d) \in (Y_\ell, X_\ell)$ について,

$$p(y, d, \theta, \boldsymbol{\beta}) = p(y|\theta) \prod_{v=1}^{V} p(v|\theta, \boldsymbol{\beta})^{n_{dv}} \cdot p(\theta)p(\boldsymbol{\beta}) \tag{19}$$

を最大化する $\boldsymbol{\beta}$ を求めることを考える. ここで第 1 項はロジスティック回帰

$$p(y=1|\theta) = \sigma(\theta) = \frac{1}{1 + e^{-\theta}} \tag{20}$$

である. このグラフィカルモデルを図 6 に示した. θ が大きい, あるいは小さい方が第 1 項の教師データに対する識別モデルの尤度が高くなるが, 逆に第 2 項の単語の生成確率が下がる可能性があるため, 両者のトレードオフで θ が決まり, それによって式 (16) から $\boldsymbol{\beta}$ が決まることになる.

ここで問題なのは, 回帰パラメータ $\boldsymbol{\beta}$ だけでなく, テキストの潜在的な極性 θ も未知なことである. θ および $\boldsymbol{\beta}$ を同時に最適化することも可能であるが, 心理統計学においてこうした同時推定は一致性を持たないことが知られている [18]. また, 二値ラベルの y という弱い教師情報からは θ は一意には決まらず, θ を学習時に点推定することは教師データへの過学習をもたらす可能性がある.

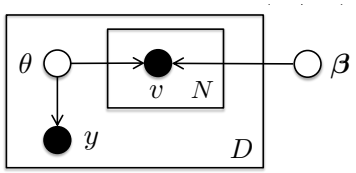


図 6 $\boldsymbol{\beta}$ を推定するための教師データのグラフィカルモデル.

[*6] この方法はうまく働くが, より精密な半教師あり学習の理論に基づいて式 (18) の目的関数を置き換えることは今後の課題としたい.

適応的ガウス–エルミート求積による解法　そこで, θ を推定する代わりにモデルから積分消去し,

$$p(y,d,\boldsymbol{\beta}) = \int_{-\infty}^{\infty} p(y,d,\theta,\boldsymbol{\beta})d\theta \tag{21}$$

$$= \int_{-\infty}^{\infty} p(y|\theta) \prod_{v=1}^{V} p(v|\theta,\boldsymbol{\beta})^{n_{dv}}\, p(\theta)d\theta \cdot p(\boldsymbol{\beta}) \tag{22}$$

を $\boldsymbol{\beta}$ について最適化することを考える.[*7] $p(\theta)$ は標準正規分布 $\mathcal{N}(0,1)$ であるから, 式 (22) の形のガウス分布に関する積分はガウス–エルミート求積と呼ばれる方法で, きわめて正確に数値的に求めることができる.

ガウス–エルミート求積では, 関数空間での直交多項式を用いて, 関数 $f(x)$ の e^{-x^2} に関する積分を次の形で高精度に近似する.

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx \simeq \sum_{i=1}^{H} w_i f(x_i) \tag{23}$$

ここで $\mathbf{x}=(x_1,\cdots,x_H)$ は分点 (abscissa) と呼ばれる座標, $\mathbf{w}=(w_1,\cdots,w_H)$ は対応する重みであり, $H=7$ のとき (本研究では $H=20$ とした) は

$$\begin{aligned}\mathbf{x} &= (-2.652, -1.674, -0.816,\ 0,\ 0.816,\ 1.674,\ 2.652),\\ \mathbf{w} &= (0.001,\ 0.055,\ 0.426,\ 0.810,\ 0.426,\ 0.055,\ 0.001)\end{aligned}$$

である. これらの値は, 標準的なガウス–エルミート求積のパッケージで計算することができる.[*8] 式 (23) は e^{-x^2} についての積分なので, $e^{-x^2}=e^{-\theta^2/2}$ すなわち $\theta=\sqrt{2}x$ とおけば, 変数変換により

$$\int_{-\infty}^{\infty} f(\theta)\mathcal{N}(\theta|0,1)d\theta = \int_{-\infty}^{\infty} f(\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{\theta^2}{2}}d\theta \tag{24}$$

$$\simeq \frac{1}{\sqrt{\pi}}\sum_{i=1}^{H} w_i f(\sqrt{2}x_i) \tag{25}$$

と計算することができる.

ただし, 式 (25) による積分は, そのままでは非常に効率が悪い. 一般にテキストは多くの単語を含むため, θ の事後分布はある値 $\hat{\theta}$ の近くに集中しており, 式 (25) ではほとんどの分点での尤度が 0 になってしまうからである. そこで, [20] の方法を用いて, 積分を $\hat{\theta}$ の周りで実行することにする. θ の事後分布を近似する平均 $\mu=\hat{\theta}$ と分散 σ^2 は二分探索と 2 階差分により容易に求めることができるので, まず, $\phi=\mu+\sigma\theta$ と変数変換すると, 簡単な計算により

$$\int_{-\infty}^{\infty} f(\theta)\mathcal{N}(\theta|\mu,\sigma^2)d\theta \simeq \frac{1}{\sqrt{\pi}}\sum_{i=1}^{H} w_i f(\mu+\sqrt{2}\sigma x_i) \tag{26}$$

であることがわかる. このとき, 求める積分を次のように変形する.

$$I = \int_{-\infty}^{\infty} f(\theta)\mathcal{N}(\theta|0,1)d\theta \tag{27}$$

$$= \int_{-\infty}^{\infty} \underbrace{f(\theta)\frac{\mathcal{N}(\theta|0,1)}{\mathcal{N}(\theta|\mu,\sigma^2)}}_{h(\theta)}\mathcal{N}(\theta|\mu,\sigma^2)d\theta \tag{28}$$

式 (28) の最初の 2 項を $h(\theta)$ とおけば, 式 (26) より

$$I \simeq \frac{1}{\sqrt{\pi}}\sum_{i=1}^{H} w_i h(\mu+\sqrt{2}\sigma x_i) \tag{29}$$

と, $\mathcal{N}(\theta|\mu,\sigma^2)$ についての適応的な積分で置き換えて求めることができる.

われわれの場合, 求めたい積分は式 (22) だったから, 対数で計算するために

$$\ell(\theta) = \log p(y|\theta) + \sum_{v=1}^{V} n_{dv}\log p(v|\theta,\boldsymbol{\beta}) \tag{30}$$

と定義すれば, $h(\theta)$ は

$$h(\theta) = e^{\ell(\theta)}\frac{\mathcal{N}(\theta|0,1)}{\mathcal{N}(\theta|\mu,\sigma^2)} \tag{31}$$

$$= \exp(\ell(\theta) + \log\mathcal{N}(\theta|0,1) - \log\mathcal{N}(\theta|\mu,\sigma^2)) \tag{32}$$

となる. 見やすくするために $y_i=\mu+\sqrt{2}\sigma x_i$ とおけば,

$$p(y,d,\boldsymbol{\beta}) = \int_{-\infty}^{\infty} h(\theta)\mathcal{N}(\theta|\mu,\sigma^2)d\theta \cdot p(\boldsymbol{\beta}) \tag{33}$$

$$\simeq \frac{\sigma}{\sqrt{\pi}}\sum_{i=1}^{H}\exp\left[\log w_i + \ell(y_i) + \frac{1}{2}\left(\frac{1}{\sigma^2}(y_i-\mu)^2 - y_i^2\right)\right] \cdot p(\boldsymbol{\beta}) \tag{34}$$

となり, この対数を $\boldsymbol{\beta}$ について偏微分し, L-BFGS 法で最適化することで $\boldsymbol{\beta}$ を計算する.

4. 関連テキストの確率的抽出

3 章の方法でテキストの潜在的なスケール θ が推定できるようになったが, 分析の対象となるテキストがアンケートの場合のように事前に定まっていない場合は, 対象となるテキストを最初にコーパスから抽出する必要がある. 一般に, 分析に使用される新聞記事や SNS のテキストといったコーパスは, 分析したい対象とまったく関係のない多種多様な内容を含んでおり, その量も厖大だからである. 例えば, 米大統領のスピーチのうち「中東経済」に関係するのはそのごく一部分であるし,「エアコン」を含む Twitter のツイートは, 筆者の手元にある Twitter 公開ストリームから得た 100 万件の日本語ツイートのうち 67 件 (0.007%) にすぎず, ほとんどは全く関係のないツイートであるといってよい.

こうした目的のための方法として, 政治学方法論の分野では (1) 潜在トピックモデルを学習し, 分析対象に関連するトピックを人手で選択する, あるいは (2) キーワードでコーパスを検索し, キーワードが含まれるテキストを分析対象とするというアプローチが取られてきた. しかし, トピックを選択する方法は,「北アフリカへの経済援助」の

[*7] Hamiltonian MCMC 法 [19] を用いた $\boldsymbol{\beta}$ のベイズ推定も検討したが, MAP 推定を用いる方が安定した結果となった.

[*8] Python では `numpy.polynomial.hermite.hermgauss` で, R では `gaussquad` パッケージの `hermite.h.quadrature.rules` で計算できる.

ような複数のトピックにまたがる話題を扱うことができないという問題がある. たとえ「北アフリカ」と「経済援助」にそれぞれ対応するトピックが同定できたとしても, その重みを同等としてよいかは自明ではない. また, 分析に使われるコーパスが大規模化し, 特にトピックモデルに使われるトピック数が容易に 1000 を超えるような場合 [21] は, 人手でトピックを同定することは困難であり, その結果の客観性も保証されない.

一方でキーワードで検索する方法は客観的であるが, たとえ意味的に近くても, そのキーワードを含まない場合にテキストを取り出せない, という問題がある. たとえば「衆議院議員選挙」に関するツイートを検索する方法では,「衆院選」や「国政選挙」といった言葉を使ったツイートは検索に漏れてしまう. こうした同義語をどの程度まで検索に含めるかに客観的な指針はなく[*9], 人手では漏れが生じる可能性が非常に高い.

LSS で用いられているように, 共起頻度や単語ベクトルを用いて, 関連度を表す何らかのスコアを定義してその上位を用いることでこれらの問題は解決できるが, こうしたヒューリスティックな方法では, スコアの「上位」とは何かの基準が分析対象によって毎回異なってしまい, 客観的な定義が難しいという問題がある. このため, こうした方法は社会科学に特に求められる客観性や再現性 (reproducible research [22]) に繋がらない.

そこで本研究では, (1) 潜在トピックモデルおよび (2) ニューラル文書ベクトルを用いて, テキストの, キーワード集合 T への関連度を確率として求める方法を提案する. これらは人手やヒューリスティックに依存せず, 確率として得られるために確率の閾値を変えることで, 分析に用いるテキストの粒度を客観的に調整することができる. これらの方法は本稿で提案する PLSS だけでなく, Wordfish 等の他の分析にも同様に用いることができる, 基礎的な方法である.

4.1 潜在トピックモデルに基づく方法

確率的な潜在トピックモデル [7] は, 政治学におけるテキスト分析においても標準的に使われる方法となっている [2][23]. いま, LSS における「目標語」, すなわち分析の対象とするキーワード集合 T が $T=\{$“労働”, “賃金”$\}$ のように与えられたとしよう. このとき, コーパスの各テキストが T と関連する確率をどのように求めたらいいのだろうか.

直感的には, テキストに含まれる単語のうち T と関連する単語の割合が高いほど, そのテキストは T と関連していると考えられる. しかし, テキストには英語ならば “is”, “the”, “thing”, “way” など内容と関連の薄い背景語 (機能語) が多く存在し, それらは分野によるために事前にリスト化することはできない. 例えば, 論文であれば「研究」や「モデル」はそれ自体ではほぼ意味を持たない単語であるし [24], こうした単語は分野ごとに無数に存在する.

そこで, こうした背景語が専用の背景分布 $p_0(\cdot)$ から生成されたとする, 背景付き LDA[24][25] を考える. この背景付き LDA(以下 LDAb と呼ぶ[*10]) では, 各テキスト d の単語 w_{dn} は次のように生成されたと仮定する. ここで Dir はディリクレ分布, Be はベータ分布を表す.

Draw $p_0(\cdot) \sim \mathrm{Dir}(\eta)$.
For $k = 1 \cdots K$, Draw $p(\cdot|k) \sim \mathrm{Dir}(\beta)$.
for $d = 1 \cdots D$ **do**
  Draw $\lambda \sim \mathrm{Be}(a, b)$; Draw $\theta \sim \mathrm{Dir}(\alpha)$.
  for $n = 1 \cdots N_i$ **do**
    if $\mathrm{Bernoulli}(\lambda)$ **then**
      Draw $w_{dn} \sim p_0(\cdot)$
    else
      Draw $z_{dn} \sim \theta$
      Draw $w_{dn} \sim p(\cdot|z_{dn})$.
    end if
  end for
end for

すなわち, このモデルでは各単語はテキストごとに異なる背景確率 λ で背景分布 $p_0(\cdot)$ から生成され, そうでなければ通常のトピックモデルと同様にテキストのトピック分布に従って生成されたと考える. ここで, 最後のステップ 9 と 10 は z_{dn} について期待値をとれば, $\boldsymbol{\beta}=\{p(\cdot|k)\}_{k=1}^{K}$ を混合比 θ で混ぜ合わせたユニグラム分布

$$w_{dn} \sim \sum_{k=1}^{K} \theta_k p(\cdot|k) = \mathrm{Mixture}(\boldsymbol{\beta}, \theta) \qquad (35)$$

から 1 ステップで生成されたと考えることができるから, この分布を右辺のように, $\mathrm{Mixture}(\boldsymbol{\beta}, \theta)$ と以後表記する.

背景分布 $p_0(\cdot)$ はトピック分布よりごく一部の語に確率が集中していると考えられることから, $\eta < \beta$ に設定することで, ギブスサンプリングにより, 上の背景付き LDA を与えられたコーパスについて学習することができる. LSS に付属する, 英国の新聞 The Guardian から採られた 10,000 記事からなる Guardian コーパスおよび, 日本語の Livedoor

表 2 Guardian (左) および Livedoor (右) コーパスで学習された背景分布 $p_0(\cdot)$ の上位語. 一般的なストップワードに含まれない単語も, 統計的に背景語と学習されていることがわかる.

the	0.1351	was	0.0155	の	0.1038	も	0.0239
to	0.0656	said	0.0138	、	0.0678	な	0.0181
of	0.0571	with	0.0136	に	0.0660	する	0.0136
a	0.0511	as	0.0136	を	0.0636	いる	0.0135
and	0.0506	be	0.0121	が	0.0583	ない	0.0130
in	0.0473	at	0.0119	は	0.0562	こと	0.0123
that	0.0272	by	0.0116	て	0.0509	さ	0.0114
’s	0.0238	have	0.0113	。	0.0489	だ	0.0110
is	0.0217	has	0.0110	で	0.0424	から	0.0109
for	0.0217	are	0.0109	た	0.0421	れ	0.0100
on	0.0207	from	0.0102	と	0.0385	い	0.0090
it	0.0182	not	0.0096	し	0.0313	か	0.0088

[*9] 本稿で提案するような統計的な方法に基づき, いかにしてキーワード検索を自動的に発行し, その結果を統合するかは興味深い問題である. たとえ高速な方法を用いても, 巨大な SNS テキスト全体に対して統計モデルを計算することは現実的ではないからである.

[*10] http://chasen.org/~daiti-m/dist/ldab/ でテキストから直接学習できる Cython 実装を公開している.

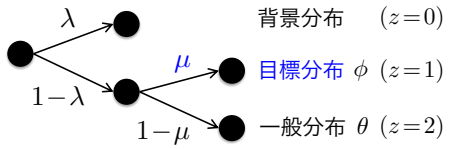


図 7 背景トピックモデルを用いたテキストの関連確率の推定. テキストの各単語は背景分布, 目標分布, 一般分布のいずれかに属し, 目標分布からキーワード集合 T が生成されたと仮定する.

ニュースコーパス[*11]について学習された背景分布 $p_0(\cdot)$ の例を表 2 に示した. このモデルでは, ステップ 4 からわかるように, テキスト毎に背景語の割合 λ が違っていてよく, たとえば子供の書いた言葉のように背景語の割合が高いテキストが含まれていても, それを考慮して $p_0(\cdot)$ およびトピック分布 $\boldsymbol{\beta}$ を学習することができる.

キーワード集合 T との同時生成モデル この上で, キーワード集合 T が与えられたとき, コーパスの各テキストについて図 7 に示したように, 次の仮想的な生成モデルを考える. 以下, 文書インデックスを表す d を省略する.

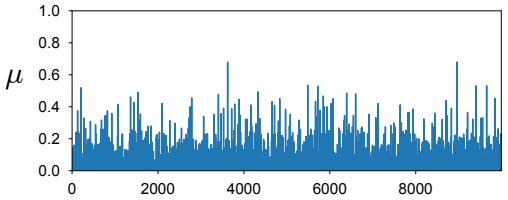
1: Draw $\lambda \sim \mathrm{Be}(a_0, b_0)$; $\mu \sim \mathrm{Be}(a_1, b_1)$; $\theta \sim \mathrm{Dir}(\alpha)$.
2: **for** $n = 1 \cdots N$ **do**
3: **if** Bernoulli(λ) **then**
4: Draw $w_n \sim p_0(\cdot)$.
5: **else**
6: **if** Bernoulli(μ) **then**
7: Draw $w_n \sim \mathrm{Mixture}(\boldsymbol{\beta}, \phi)$.
8: **else**
9: Draw $w_n \sim \mathrm{Mixture}(\boldsymbol{\beta}, \theta)$.
10: **end if**
11: **end if**
12: **end for**

ここで ϕ は T が生成されたトピック分布であり, T は次のようにして生成されたと仮定する.

1: **for** v in T **do**
2: $v \sim \mathrm{Mixture}(\boldsymbol{\beta}, \phi)$.
3: **end for**

この ϕ は, 学習された背景付き LDA について変分ベイズ EM アルゴリズム [7] で高速に計算することができる. すなわち, 背景分布を除いた上で, 各単語は確率 μ で T と同じ分布 $\mathrm{Mixture}(\boldsymbol{\beta}, \phi)$ から生成され, 確率 $1-\mu$ で通常のトピック混合分布 $\mathrm{Mixture}(\boldsymbol{\beta}, \theta)$ から生成されたとする. この μ が T との関連確率を表しており, μ を推論することが目的となる.

これは学習時には図 7 右に示したように, 各単語 w にその生成された情報源を表す潜在変数 $z \in \{0, 1, 2\}$ が付与された確率モデルとみなすことができる. $z=0$ のときは w は p_0 から, $z=1$ のときは $\mathrm{Mixture}(\boldsymbol{\beta}, \phi)$ から, $z=2$ のときは $\mathrm{Mixture}(\boldsymbol{\beta}, \theta)$ から生成されたことに対応し, それぞれの分布は既知であるから[*12], これは簡単な EM アルゴリズムで推定することができる. E ステップで各単語の z の事後確率がわかれば, M ステップで μ (および λ) はベータ事後分布の期待値として計算でき, これを収束するまで反復する.



(a) 各テキストで計算された T との関連確率 μ.

```
britain confederation industry warned growth
prospects uk economy cutting forecasts cbi
weaker performance end year coupled less rosy
outlook household spending trade investment growth
downgrade expected economic growth measured gross
domestic product percentage reduction ...
```

(b) $\mu = 51.9\%$ となったテキスト中の関連単語 ($z=1$).

図 8 Guardian コーパスで $T = \{$“economy”, “money”$\}$ とした際の各テキストの関連確率 μ の計算.

図 8(a) に, Guardian コーパスで $T = \{$“economy”, “money”$\}$ とした場合の, 各テキストの T との関連確率 μ のプロットを示す. 事前に計算する背景付き LDA のトピック数は $K = 100$ とした. 大部分のテキストでは μ は 1%未満のごく小さな値となるが[*13], 一部のテキストでは 50%を超える値となり, T と関連が深いテキストであることがうかがえる. こうしたテキストの例を図 8(b) に示した.

この方法は各単語に対する統計モデルであるため, テキストがキーワード集合に関係するかだけでなく, その中で関係している単語, すなわち $z_{dn}=1$ となった単語だけを抽出することもできる. 図 9 に, Livedoor コーパスにおいて $T = \{$“美容”, “化粧”$\}$ に関係した単語だけを抽出した例を示した.[*14] このように, 本手法により教師なしで, 背景語を除いた上でさらに関連する語だけを確率的に抽出することが可能になる.

ただし, 政治学のような場面では, ある言葉がどのような文脈で使われているかが重要なことが多く, T に関係した単語 ($z=1$) だけを抽出するより, テキストの内容語全体 ($z=1,2$) を抽出した方がよい場合が多いと考えられる.

4.2 ニューラル文書ベクトルに基づく方法

上の背景付きトピックモデルによる関連確率の計算は精密ではあるが, トピックモデルの学習が必要なため, 計算量が非常に大きいという問題がある. 一般にテキスト抽出

春 磨く エステ 美人 陽気 春 明るい メイク 服装 冬 ケア 体 フェイス ライン足 カサカサ 乾燥 ！ 対処 たく 顔 ボディ 足 全身 エステ ケア しまい ます そこで 気 箇所 ケア エステ 紹介 簡単 自宅 ケア もらい 口 エステ フェイス ライン スッキリ ！ エッセンス エステ 集め 口 内側 マッサージ 口 筋肉 刺激 皮膚 届きにくい マッサージ マッサージ 顔 筋肉 コリ 顔 や フェイス ライン むくみ 解消 ます ...

図 9 Livedoor コーパスにおいて $T = \{$“美容”, “化粧”$\}$ として関連語のみ ($z=1$) を抽出したテキスト. このテキストの T との関連確率は $\mu = 45.83\%$ であった.

[*11] https://www.rondhuit.com/download.html#ldcc
[*12] θ は未知であるため, $\mathrm{Mixture}(\boldsymbol{\beta}, \theta)$ では θ に対する期待値をとる, すなわち $\theta_k = 1/K$ として学習を行う.
[*13] LDA では $\boldsymbol{\beta}$ にディリクレスムージングが使われているため, 全く関係のないテキストでも, こうしたごく小さな確率が発生する.
[*14] EM アルゴリズムで期待値を計算しているため, $p(z=1|w) > 0.5$ となった単語だけを示した.

前のコーパスは様々な内容を含むため膨大であり, 背景トピックモデルの学習には数万文書程度でも一晩前後の時間が必要になる.

そこで, もう一つの方法として, Word2Vec と同等の意味を持ち, 高速に計算できる文書ベクトル [8] を利用した統計モデルを提案する. Levy ら [26] が示したように, Word2Vec の Skip-gram は, 単語 w とその周辺語 c について, 次の Shifted Positive PMI

$$Y(w,c) = \max\left(\log \frac{p(w,c)}{p(w)p(c)} - \log k, 0\right) \tag{36}$$

を要素とする行列 $\mathbf{Y}$ を考え, 特異値分解により $\mathbf{Y} \simeq \mathbf{W}\mathbf{C}^T$ と行列分解したときの行列 $\mathbf{W}$ の各行を求めることと数学的に等価である. ここで k は Skip-gram での負例数に相当するが, 行列分解の場合は [27] により $k=1$ とする, すなわちシフトを行わないことが最適であることが示されており, 以下では $k=1$ を用いた.

これを文書に拡張すると, 図 10 のように単語 w と文書 d の PPMI を同様に

$$Y(w,d) = \max\left(\log \frac{p(w,d)}{p(w)p(d)} - \log k, 0\right) \tag{37}$$

$$= \max\left(\log \frac{p(w|d)}{p(w)} - \log k, 0\right) \tag{38}$$

と定義し, $\mathbf{Y} \simeq \mathbf{W}\mathbf{D}^T$ と行列分解したときの $\mathbf{W}$ および $\mathbf{D}$ の各行として単語ベクトルおよび文書ベクトルが得られる [8]. $p(w|d)$ および $p(w)$ は, 頻度から容易に計算することができる.

このとき, T に含まれる語に 1 が立ち, 他は 0 の検索ベクトルを $\mathbf{y}$ とすれば, この検索ベクトルに対応する仮想的な文書ベクトル $\mathbf{d}$ は図 10 のように, 二乗誤差の意味で

$$\mathbf{y} \simeq \mathbf{W}\mathbf{d} \tag{39}$$

の関係がある. よって, これは通常の線形回帰 (OLS) であり, $\mathbf{d}$ の最適解は

$$\mathbf{d} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{y} \tag{40}$$

と求められる. 事前に回帰行列 $\mathbf{R} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T$ を計算しておけば,

$$\mathbf{d} = \mathbf{R}\mathbf{y} \tag{41}$$

と一瞬で求めることができる.[*15]

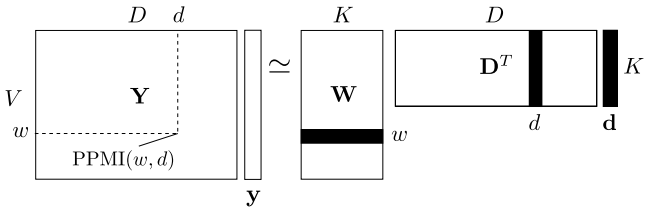


図 10 行列分解による単語ベクトルおよび文書ベクトルの計算. こうして得られる単語ベクトルは, Word2Vec によるものと数学的に等価である. このとき, 検索ベクトル $\mathbf{y}$ には仮想的な文書ベクトル $\mathbf{d}$ が対応し, OLS によって解析的に求められる.

[*15] なお, 文書ベクトルとして単純なニューラル手法として知られる DocVec[28] を使った場合はこうした解析解は存在せず, $\mathbf{d}$ を求めるには数値的最適化が必要になる. また, 学習には特異値分解に比べて 10 倍以上の時間がかかり, 性能も数学的な最適解である式 (40) に比べて落ちることを確かめている [8].

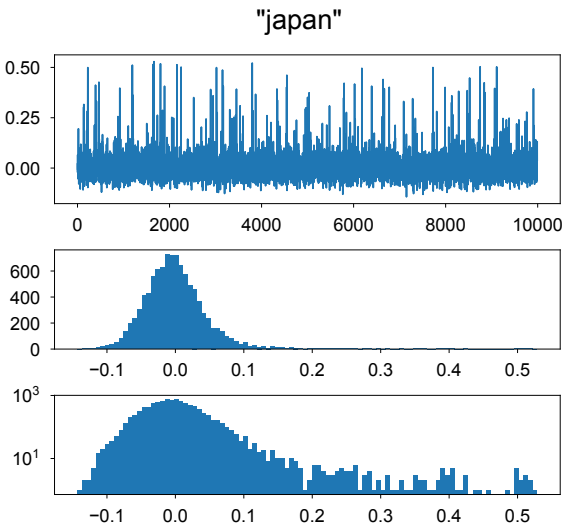


図 11 Guardian コーパスにおいて, $T=\{$"japan"$\}$ と指定したときの各テキストの類似度スコア (上段) とそのヒストグラム (中段：線形スケール, 下段：対数スケール). ヒストグラムの右側に, 線形スケールでは見えない外れ値が薄く分布している. 抽出された外れ値は, 外れ値確率の閾値を 0.5 として 118 個であった.

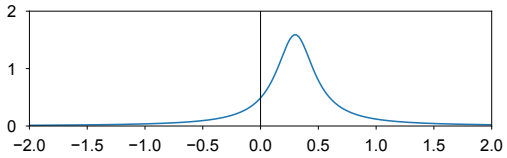


図 12 コーシー分布 Cauchy$(0.3, 0.2)$ の確率密度関数. 裾が非常に広い確率分布となっている.

キーワード集合 T に対応する文書ベクトル $\mathbf{d}$ が得られれば, これと他の文書ベクトル, すなわち $\mathbf{D}$ の各行とのコサイン類似度を $\mathbf{x} = \widetilde{\mathbf{D}}\widetilde{\mathbf{d}}$ のように計算することで, T と各テキストの類似度スコアを求めることができる. ここで, $\widetilde{\mathbf{d}}$ と $\widetilde{\mathbf{D}}$ は $\mathbf{d}$ および $\mathbf{D}$ の各行のノルムを 1 に正規化したベクトルおよび行列である. 図 11 に, $T=\{$"japan"$\}$ のときに Guardian コーパスに含まれる 10000 個のテキストについて計算したスコアの例を示した.

正規–コーシー混合モデルによる関連確率の計算 ただし, これだけからは, どのスコアまでのテキストを T に関連するとして抽出すべきかは自明ではない. ここで図 11 上段のスコアをヒストグラムにすると図 11 中段のようになっており, ほとんどのテキストに対するスコアは正規分布をなしているが, 右側に上段にみえる「外れ値」が存在していることがわかる. これは, 図 11 下段のように対数スケールで表すと特に顕著になる.

そこで, こうした「外れ値」, すなわち類似度が全体と有意に異なっているテキストを T と関連しているテキストとみなし, 図 11 中段のようなスコアの分布を正規分布および, 外れ値を表すコーシー分布の混合分布としてモデル化

する．コーシー分布は自由度 1 の t 分布であり，外れ値のモデルによく使われる確率分布で，確率密度関数は位置パラメータ c と尺度パラメータ $\gamma > 0$ を用いて

$$\mathrm{Cauchy}(x|c,\gamma) = \frac{1}{\pi}\frac{\gamma}{(x-c)^2+\gamma^2} \tag{42}$$

で表される (図 12)．これから，観測スコア $\mathbf{x}$ の生成モデルは以下のようになる．

1: Draw $\lambda \sim \mathrm{Be}(1,1)$.
2: **for** $i = 1 \cdots D$ **do**
3: **if** $\mathrm{Bernoulli}(\lambda)$ **then**
4: $x_i \sim \mathrm{Cauchy}(c,\gamma)$
5: **else**
6: $x_i \sim \mathcal{N}(\mu,\sigma^2)$
7: **end if**
8: **end for**

観測スコア集合 $\mathbf{x}$ に対するこの混合モデルは，EM アルゴリズムを用いて高速に推定できる．E ステップでは各スコア x_i がガウス分布 $\mathcal{N}(\mu,\sigma^2)$ とコーシー分布 $\mathrm{Cauchy}(c,\gamma)$ のどちらに属するかの期待値 $p(z_i|x_i)$ を計算し，M ステップではその期待値に基づいて，正規分布およびコーシー分布のパラメータを最適化する．コーシー分布のパラメータ c,γ には解析的な更新式がないため，BFGS 法により数値的に最適化する．γ は正なので $\gamma = e^t$ と置換すれば，対数尤度の微分は

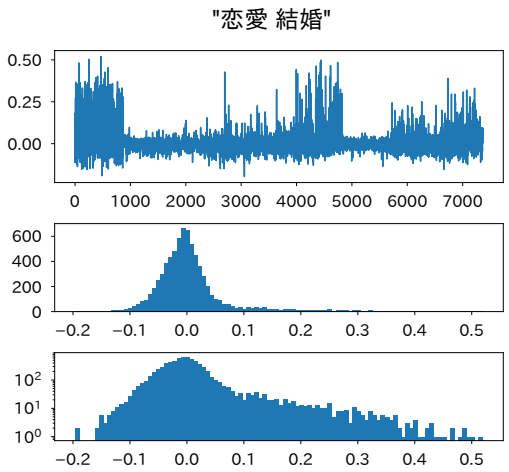
$$\begin{cases} \dfrac{\partial}{\partial c}\log \mathrm{Cauchy}(x|c,e^t) = \dfrac{2(x-c)}{(x-c)^2+e^{2t}} \\ \dfrac{\partial}{\partial t}\log \mathrm{Cauchy}(x|c,e^t) = 1 - \dfrac{2}{1+(x-c)^2e^{-2t}} \end{cases} \tag{43}$$

となり，これを BFGS 法の勾配として用いる．図 13 に，全体の推定アルゴリズムを示した．

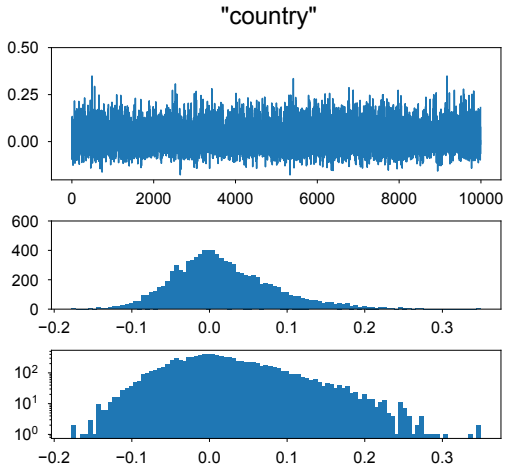
特異値分解による文書ベクトルおよび外れ値検出の EM アルゴリズムはいずれも高速であり，トピックモデルより圧倒的に速く，多くの場合その場で計算を終えることができる．図 14 および付録 B に，Guardian コーパスに対して $T = \{$“country”$\}$，Livedoor コーパスに対して

$\mu = \mathrm{median}(\mathbf{x})$, $\sigma^2 = \mathrm{var}(\mathbf{x})$
$c = \mathrm{mean}(\mathbf{x})$, $\gamma = 1$
$q_0 = 0.5$, $q_1 = 0.5$ (* From $\mathrm{Be}(1,1)$ *)
while not converged **do**
  (* **E step** *)
  for $n = 1 \cdots N$ **do**
  Compute $p_0 = q_0 \cdot \mathrm{Cauchy}(x_i|c,\gamma)$
  Compute $p_1 = q_1 \cdot \mathcal{N}(x_i|\mu,\sigma^2)$
  $Z[n] = [p_0/(p_0+p_1), p_1/(p_0+p_1)]$
  end for
  (* **M step** *)
  $\mu = \mathbf{x} \cdot Z[:,1]/\mathrm{sum}(Z[:,1])$
  $\sigma = (\mathbf{x}-c) \cdot Z[:,0]/\mathrm{sum}(Z[:,0])$
  Optimize c,γ via BFGS
  (* **H step** *)
  $q_0, q_1 = \mathrm{normalize}([1,1] + \mathrm{sum}(Z,0))$
end while

図 13 正規–コーシー混合モデルの EM アルゴリズム．



(a) Livedoor コーパスで $T = \{$“恋愛”, “結婚”$\}$ とした場合．関連するテキストは「独女通信」カテゴリ (1〜1000 付近) および「Peachy」カテゴリ (4000 付近) に多いが，「スポーツ」(6000 付近) および「トピックニュース」(7000 付近) にも存在している．確率 0.5 以上の外れ値は 262 件であった．



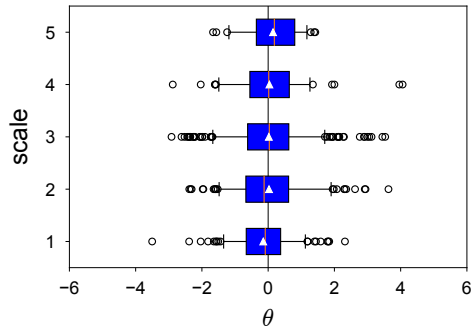
(b) Guardian コーパスで $T = \{$“country”$\}$ とした場合．“country” は一般的な語であるため，類似度が有意に異なる外れ値がほとんど存在せず，右端の 2 個のみとなった．

図 14 キーワード集合 T を与えた際の各テキストの類似度スコアとその分布．中段は線形スケール，下段は対数スケールで表している．

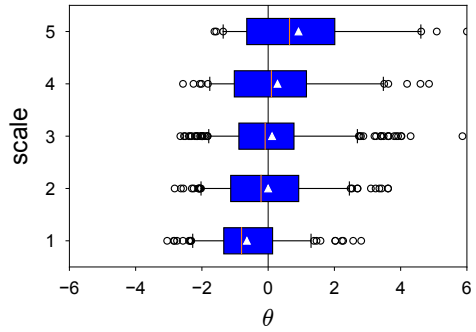
$T = \{$“恋愛”,“結婚”$\}$ を指定して抽出されたテキストおよびスコア全体 $\mathbf{x}$ の分布を示した．“country” は一般的な語であるため，類似度が有意に異なるテキストはほとんどなく，閾値を 0.5 として抽出されたのは 2 件のみであった．これに対して，“恋愛”, “結婚” はコーパス中で主要な話題に属するため，262 件のテキストが抽出されている．この方法はテキストが T に有意に関係する (=スコアが外れ値である) 確率 p を求めるため，p の閾値を調整することで，T に強く関与するテキストのみを抽出するか，より広く関係するテキストを抽出するかを統一的にユーザーが選択できるのが特徴である．

5. 実験

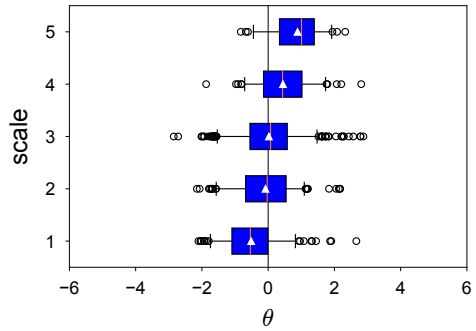
政治学の分野で公開されているテキストの極性評価デー



(a) LSS によるスコアの分布.



(b) PLSS によるスコアの分布.



(c) PLSS に半教師ありデータを加えた場合のスコアの分布.

図 15 Young and Soroka (2012) の公開データでの LSS と PLSS の結果の比較. △は各スケールでの平均を表している.

タ, および日本語・中国語の生テキストを使って実験を行った.

5.1 テキスト極性公開データでの評価

Young and Soroka [13] は, 彼らの辞書ベースの手法を評価するためのテキストとその極性評価データを公開している.[*16] これは New York Times から採られた 900 個の記事からなり, 半分は 1988–2008 年の経済面から, 残りの半分は 2007–2009 年の環境・外交・犯罪に関する一面記事からランダムにサンプリングされたものである. 各記事は専門の評価者 3 人により, 論調について Positive/Negative/Neutral のタグが付与されている. [13] と同様に, この 3 個のタグを表 3 のルールで 1–5 の 5 段階に分け, この評定と PLSS および LSS が出力する θ がどの程度相関するかを調べた. 3 人の評価が一致しないことも多いため, この「正解」の 1–5 のスケールは本質的にノイズを含んでいることに注意されたい.

図 15 に, 標準的な表 1 の極性辞書を用いて推定された θ をスケール毎に示した. 提案法は Bag of Words の単純なモデルであり, 教師データを用いた識別モデルでもないため分散は大きいが, LSS では各スケールで θ がほとんど同じ分布になっているのに対し, PLSS では明らかにスケールと θ に正の相関がみられることがわかる. ピアソンの相関係数を計算したところ, LSS では相関係数が 0.065 とほとんど 0 なのに対し, PLSS の相関係数は 0.240 となった.

LSS においても, 各スケールで平均を取れば△で示したように値はスケールと相関しており, 相関係数は 0.901 となった. LSS の論文での評価は, 常にそうして行われたものであることに注意されたい. PLSS の場合は平均はさらに高い相関を持ち, 相関係数は 0.965 となった. これから, PLSS は単純なモデルであるにもかかわらず, LSS より分散が小さく, 人手のスケールと高い相関を見せる手法であるといえる. 半教師ありデータとして, 正例と負例を 15 件ずつ加えて $\boldsymbol{\beta}$ を学習した場合 (図 15(c)) は相関係数は 0.985 とさらに相関が高くなり, 文書ごとの θ も分散が小さくなって, 相関係数も 0.348 と上昇することがわかった. 否定や引用など, 極性に関わるテキスト内部の構造は現在は考慮していないため, これらを考慮することで, 精度はさらに高くなると予想される.

5.2 生テキストの分析

5.2.1 中国語テキストの分析

御器谷 [29] は, LSS を用いて中国共産党の機関誌である『人民日報』のテキストを 1974 年から 1994 年にかけて分析し, 「伝統 (传统)」をキーワードに抽出した 39,297 件のテキストに LSS を適用することで, 中国共産党において「伝統」概念が当初否定的に評価されていたものの, 党のプロパガンダに使われる中で次第に肯定的に使われていく様子を示した. ここでは極性辞書として, 表 4 に示した中国語の標準的な肯定的/否定的単語のリストを用いている.

このデータを同様に PLSS で分析したところ, 図 16 のように大きな傾向は同様であるが, より動きの大きな結果が得られた. このとき, PLSS と LSS でそれぞれ計算された単語の極性を表 5 および表 6 に示した. LSS においては極性の大きい単語は一般的な肯定語・否定語となっているが, PLSS では全体に極性のレンジが広がっており, 最も極性が大きいのは中国共産党という文脈で党のために肯定さ

表 3 3 人分の positive/negative/neutral の評価データから 5 段階のスケールへの変換ルール. ほぼ [13] のものと同様である.

条件	スケール
#(positive) = 3	5
#(positive) = 2	4
それ以外	3
#(negative) = 2	2
#(negative) = 3	1

表 4 中国語の正負の標準的な極性語辞書.

+	良好, 积极, 完美, 善良, 优良, 正确, 理想
−	不好, 缺乏, 消极, 贪心, 固执, 反动, 份子

[*16] このデータは, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/H2FEPO から入手することができる.

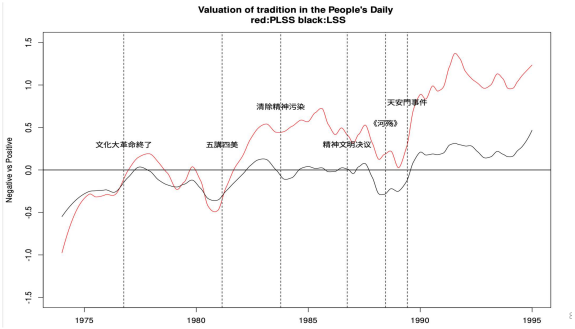


図 16 LSS (黒) と PLSS (赤) による,『人民日報』の「伝統」を含む記事の極性の時間的変化. 上下が正負の極性に対応している.

表 5 PLSS で計算された『人民日報』の単語の極性.

v	ϕ_v	意味
振兴中华	+0.5716	中華を振興する
优异成绩	+0.5289	優れた成績
和衷共济	+0.5271	心を合わせて助け合う
良好	+0.5148	良い、素晴らしい
大业	+0.5058	偉大な事業
中日	+0.4898	中国と日本(中日関係)
祖国统一	+0.4882	祖国の統一
努力	+0.4865	努力
统一祖国	+0.4851	祖国を統一する
残疾人	+0.4820	体の不自由な方
典范	+0.4817	手本，模範
远见卓识	+0.4814	先見の明がある

(a) $\phi_v > 0$ の上位語

v	ϕ_v	意味
帽子	−0.6409	レッテル
不敢	−0.6402	～する勇気がない
坏人	−0.6361	悪人
泛滥	−0.6342	氾濫する、はびこる
不清	−0.6318	ぼんやりとする
心有余悸	−0.6226	思い出しただけでびくびくする
散布	−0.6079	まき散らす，振りまく
帮派	−0.6069	(政治目的で結成された) 派閥
横加	−0.6011	横暴に，理不尽に，乱暴に
把持	−0.5956	(権力・組織などを)独り占めにする
乘机	−0.5879	すきをつく
打砸抢	−0.5813	(文革中に)打ち壊しを働く人々

(b) $\phi_v < 0$ の下位語

れる単語, あるいは否定される単語となっていることがわかる. LSS の場合のような一般的な極性語も存在するが, 表 5 のように, それらが最も大きな極性を持つわけではなく, PLSS の方が『人民日報』という文脈での肯定–否定の軸を捉えていることが読み取れる. なお, 図 16 において 1980 年代に極性に差があるのは, 舞台や小説など芸能面での記述が, ϕ_v の違いによって大きな極性を持つことが一因であることがわかった.

5.2.2 日本語テキストの分析

ここまでの分析では表 1 や表 4 のような極性辞書を用いていたが, たとえ少量でも, こうした辞書が分析のために常に作成可能であるとは限らない. 肯定的–否定的といった単純な軸ではなく, より複雑な意味的な軸を問題にしている場合は, どの単語が極性を決めるかが事前には明らかではないからである.

しかし, そうした場合でも 3.2 節に述べたように, 分析したい軸を特徴づける代表的なテキストを選ぶことはできる場合が多いと考えられる. そこで, 2021 年に行われた通常

表 6 LSS で計算された『人民日報』の単語の極性.

v	ϕ_v	意味
优良	+0.1873	優良な
完美	+0.1796	完璧な、すばらしい
作风	+0.1754	作風、気風
良好	+0.1669	良好な
实话	+0.1662	本当の話
高尚	+0.1655	高尚な
发扬	+0.1619	発揚する
大鼓	+0.1555	三弦の伴奏で語る語り物
不枉	+0.1507	むだではない
品质	+0.1502	本質，品性
光大	+0.1483	輝かせる，勢いよくする
俗流	+0.1480	習俗

(a) $\phi_v > 0$ の上位語

v	ϕ_v	意味
缺乏	−0.2060	～に欠ける
不好	−0.1824	悪い
瘦瘦	−0.1766	やせた
不大	−0.1709	小さい
不够	−0.1701	足りない
甚至	−0.1695	甚だしきにいたっては
信誓旦旦	−0.1686	誠意を持って本心から誓う
种种	−0.1601	様々の
不得不	−0.1592	～せざるを得ない
个别	−0.1571	ごく少数の、個々の
告密	−0.1553	密告する，告発する
不愿	−0.1552	嫌う、望まない

(b) $\phi_v < 0$ の下位語

国会の議事録を用いて, 議員を特徴づける極性軸と, それに基づく各発言の極性を計算することを試みた. ここでは議員の立場が比較的明らかになると考えられる, 衆議院の農林水産委員会の議事録を使用した.[*17] これは 12 回の開催分, 合計で 1,324 個の発言からなる.

議長および農林水産大臣の属する自民党と他の党との違いは比較的明らかであるため, ここでは発言数の多い国民民主党の玉木雄一郎氏を + 軸, 日本共産党の田村貴昭氏を – 軸として分析を行った. 両氏の発言からランダムに 20 件を選択して半教師ありデータ X_ℓ とし, 玉木氏の発言の Y_ℓ を 1, 田村氏の発言を 0 として 3.2 節の方法により $\boldsymbol{\beta}$ を最適化した. 極性辞書はまったく用いていない. この計算は式 (16) の PLSS の確率モデルに基づいているため, 単語の平均的な確率を考慮しており, また単語は常にベクトルとして意味を考慮するため, ナイーブに単語の頻度を数える方法とは全く異なることに注意されたい. 単語ベクトルとしては, 毎日新聞の 2011 年の全文 2180 万語から Word2Vec で計算した 100 次元の単語埋め込みを用いた.

表 7 に, 得られた $\boldsymbol{\beta}$ から式 (12) によって計算した単語の極性の例を示した. これから, 両氏とも野党 (左翼) であるものの, 共産党の田村氏は訴訟や裁判といった法的手続きにフォーカスしており, 一方で玉木氏は ϕ_v の上位語にみられるように, 大蔵官僚であった経験をもとに前向きに政策を実行するよう議論していることが読みとれる.

こうして得られた極性軸に沿って, 他の議員の発言の極性 θ を計算することもできる. 付録 A に, 全発言 1,324 件からランダムに選択した 40 件の発言とその θ を示した. 二

[*17] 国会の議事録は `https://kokkai.ndl.go.jp/` の国会会議録検索システムから, 委員会を指定してテキスト形式でダウンロードできる.

表 7 衆議院農林水産委員会において, 玉木 (国民民主党)–田村 (日本共産党) の発言を軸にして計算された各単語の極性 ϕ_v.

v	ϕ_v	v	ϕ_v
まずは	0.5167	訴訟	−0.5769
なので	0.4560	傍聴	−0.5646
一つ	0.4364	毀損	−0.5519
ミリ	0.4246	原告	−0.5481
増やす	0.4178	敗訴	−0.5147
整い	0.4109	控訴	−0.5010
もっと	0.4014	係争	−0.4963
もう少し	0.4012	裁判所	−0.4962
植え付ける	0.3995	判決	−0.4808
切り替える	0.3985	審	−0.4727
一番	0.3982	シベリア	−0.4599
どうにか	0.3940	退け	−0.4491
しっかり	0.3903	高裁	−0.4462
同時に	0.3897	弁護	−0.4431
早く	0.3890	最高裁	−0.4410
戦える	0.3886	紛争	−0.4303
重点	0.3788	賠償	−0.4229
とにかく	0.3769	在住	−0.4199
歩	0.3762	証言	−0.4196
試し	0.3681	棄却	−0.4105

氏の発言だけでなく, 他の委員の発言もこの軸上に連続的に位置づけられており, 各発言の性格が二氏を結ぶ極性軸上で客観的に明らかになっているといえる.

6. まとめと展望

本研究では, テキストをある意味的尺度で連続的に測定するために, 項目反応理論に基づいて LSS (潜在意味スケーリング) を確率的に拡張する形で, PLSS (確率的潜在意味スケーリング) を提案した. PLSS は LSS と比べてヒューリスティックな処理やパラメータがなく, 政治学の分野で公開されている評価データを用いて, LSS より優れた人間の評価との相関を持つことを確かめた. 確率的に定式化することで, 潜在的スケールの分散が考慮できるとともに, 欠損値への対応や時系列拡張といったモデルの高度化が見通しよく可能になる.

また, 分析に用いるテキストを選ぶ基準として背景付き潜在トピックモデルによる方法およびニューラル文書ベクトルによる方法を提案し, キーワードに意味的に関連するテキストを, 確率付きで選択することを可能にした. 特に後者は線形代数によるため高速であり, これまで発見的に行われてきた分析テキストの選択を, より意味的かつ客観的に行えるようになった. この方法は PLSS だけでなく, テキストを事前に選択することが決定的に重要な Wordfish をはじめ, 広く用いることのできる一般的な統計的方法である. 本研究では数値的評価として LSS と同じ評価データを用いたが, クラウドソーシングを用いたより大規模なデータ [30] でも検証する予定である. 提案法は基本的に教師なし学習であるため「正解」は存在しないが, テキスト選択が本来選ばれるべきテキストを落としていないかなど, その精度を客観的に検証する方法を考えたい.

教師なしのテキスト分析は自然言語処理だけでなく, 社会科学・人文科学において重要性を増している. 統計的に見通しのよい方法を開発することで, 人手のハイパーパラメータに依存しない客観性を保ちつつ, さらに高度な分析を行えるようにしたい.

謝辞

政治学でのテキスト分析の国際会議 POLTEXT 2019 にお呼びいただき, 本研究を行うきっかけとなった渡辺耕平氏 (ラザード・アセット・マネジメント), および議論していただいた福元健太郎氏 (学習院大学) と東北大学乾研究室の皆様, 実験に協力していただいた御器谷裕樹氏 (慶應義塾大学大学院政治学専攻博士課程) に感謝いたします.

参考文献

[1] Sean Gerrish and David M. Blei. Predicting Legislative Roll Calls from Text. In *ICML 2011*, pages 489–496, 2011.

[2] Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. Tea Party in the House: A Hierarchical Ideal Point Topic Model and Its Application to Republican Legislators in the 112th Congress. In *ACL 2015*, pages 1438–1448, 2015.

[3] Bianca Zadrozny and Charles Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *KDD 2002*, pages 694–699, 2002.

[4] Jonathan B. Slapin and Sven-Oliver Proksch. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, 2008.

[5] Kohei Watanabe. The Latent Semantic Scaling: Automated dictionary making technique for document scaling. `https://blog.koheiw.net/wp-content/uploads/2015/06/LSS-02.pdf`, 2015.

[6] Kohei Watanabe. Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures*, pages 1–23, 2020.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] 持橋大地. Researcher2Vec: ニューラル線形モデルによる自然言語処理研究者の可視化と推薦. 言語処理学会第 27 回年次大会 B2-2, 2021.

[9] 筒井貴士, 我満拓弥, 大城卓, 菅原晃平, 永井隆広, 渋木英潔, 木村泰知, 森辰則. 地方議会会議録コーパスの構築および政治情報システム構築を目標としたアノテーションの一提案. 自然言語処理, 21(2):125–155, 2014.

[10] Carl Edward Rasmussen and Christopher K. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[11] J. F. C. Kingman. *Poisson Processes*. Oxford Studies in Probability. Oxford University Press, 1992.

[12] S. Deerwester, Susan T. Dumais, and George W. Furnas. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[13] Lori Young and Stuart Soroka. Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231, 2012.

[14] 豊田秀樹. 項目反応理論・理論編—テストの数理—. 統計ライブラリー. 朝倉書店, 2005.

[15] Peter D. Turney and Michael L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21:315–346, 2003.

[16] Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. Ultradense Word Embeddings by Orthogonal Transformation. In *NAACL 2016*, pages 767–777, 2016.
[17] Philipp Dufter and Hinrich Schütze. Analytical Methods for Interpretable Ultradense Word Embeddings. In *EMNLP-IJCNLP 2019*, pages 1185–1191, 2019.
[18] Malay Ghosh. Inconsistent maximum likelihood estimators for the Rasch model. *Statistics & Probability Letters*, 23(2):165–170, 1995.
[19] Radford M. Neal. *MCMC Using Hamiltonian Dynamics*. Chapman and Hall/CRC, 2011.
[20] Qing Liu and Donald A. Pierce. A Note on Gauss-Hermite Quadrature. *Biometrika*, 81(3):624–629, 1994.
[21] Jinhui Yuan et al. LightLDA: Big Topic Models on Modest Computer Clusters. In *WWW 2015*, pages 1351–1361, 2015.
[22] Roger D. Peng. Reproducible Research in Computational Science. *Science*, 334:1226–1227, 2011.
[23] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082, 2014.
[24] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *NIPS 2006*, pages 241–248, 2006.
[25] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Modeling Social Annotation Data with Content Relevance using a Topic Model. In *NIPS 2009*, pages 835–843, 2009.
[26] Omer Levy and Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185, 2014.
[27] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.
[28] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *ICML 2014*, pages 1188–1196, 2014.
[29] 御器谷裕樹. 1980 年代の中国共産党による宣伝政策における「伝統」の位置. 日本選挙学会, 2021.
[30] Kenneth Benoit, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. Crowd-Sourced Text Analysis: Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review*, 110(2):278–295, 2016.

付録 A

2021 年の第 204 回通常国会・衆議院・農林水産委員会による発言を, 玉木雄一郎 (国民民主党)–田村貴昭 (日本共産党) の発言を θ の極性軸にして計算した各発言の潜在的な θ と発言内容. 1324 件の発言のうち, ランダムに 40 個を選択して表示している. $\boldsymbol{\beta}$ の計算には, 玉木・田村の両氏の発言から, ランダムに 20 件を半教師ありデータとして用いた.

θ	発言者	発言内容
1.6410	大串（博）委員	立憲民主党・無所属の大串です。　早速質疑に入ります。　貯保法ですけれども、私は、
1.4988	矢上委員	時間の関係で次の質問に移らせてもらいますけれども、低コスト化対策ですね。二問あっ
1.4838	重徳委員	だから、農水省に何の非もないのかと言っているんですよ。　例えば、大臣、富山県の御
1.3139	大串（博）委員	全くちぐはぐですね。一時的な要因で余っているんだったら、一時的に市場から切り離さ
1.1895	本郷政府参考人	木材流通に関してでございます。需給のミスマッチを起こさないように、生産、加工の事
1.0874	玉木委員	国民民主党の玉木雄一郎です。　本法案についてまず質問いたします。　先ほどから、規
1.0784	玉木委員	コロナにはいろいろなことを教えてもらったなと思ったんですが、例えば、マスク一つ国
1.0716	近藤（和）委員	石川県能登半島の近藤和也でございます。よろしくお願いいたします。　ＣＯＶＩＤ―1
1.0316	金子（恵）委員	今、イノベーションの話もされたので、済みません、順番を変えて、林業労働力の育成、
0.8315	神谷（裕）委員	そうしますと、遡れる限り遡るということだと思うんですが、そこで、先ほど議論になっ
0.7655	光吉政府参考人	お答えいたします。　農協の株式会社化につきましては、株式会社になったときに、例え
0.7329	野上国務大臣	新型コロナウイルスの状況に対して、様々なお立場の方が非常に厳しい状況に直面してい
0.6108	重徳委員	野上大臣、よく聞いていてくださいね。あほみたいな話ですわ、これ。八丁組合、元祖が
0.4997	玉木委員	国民民主党の玉木雄一郎です。　ちょっと順番を変えて、まず、備蓄米の子供食堂、子供
0.4790	青山（大）委員	繰り返しですけれども、日本一のレンコンの産地である霞ケ浦周辺、茨城県において、以
0.3310	葉梨副大臣	令和元年七月でございますけれども、日本国内で、ツマジロクサヨトウという害虫ですね
0.0967	水田政府参考人	お答えいたします。　委員御指摘の、まず空舎延長事業でございますけれども、これは、
0.0789	緑川委員	それと同様ということで、何が違うのか、私からお話ししたいと思います。　三年前の豪
0.0540	野上国務大臣	今御紹介いただきましたみどりの食料システム戦略でありますが、先般、中間取りまとめ
−0.1532	水田政府参考人	お答えいたします。　高収益作物次期作支援交付金の件について御質問いただきました。
−0.1984	本郷政府参考人	お答えをいたします。　本法案においては、成長に優れた苗木を積極的に用いた再造林を
−0.2142	神谷（裕）委員	仮にですけれども、当初、倫理の問題の中で、議員、あるいは元大臣、大臣、そういった
−0.3023	野上国務大臣	今御指摘をいただきました件につきまして、農林漁業やあるいは食品産業の分野におきま
−0.3433	水田政府参考人	お答えいたします。　新制度では、利用に関する基準を遵守することで、構造に関する技
−0.3503	八木参考人	お答えさせていただきます。　基幹的農業従事者数の減少が見込まれる現状におきまして
−0.4448	野上国務大臣	近年、農林水産、食品産業の分野におきましては、輸出の促進ですとかあるいはスマート
−0.4626	野上国務大臣	お話がありましたとおり、本年四月に、兵庫県農業会議と兵庫県の農地バンクでありますす
−0.5006	森政府参考人	お答えいたします。　委員御指摘のとおり、協議の詳細につきましてはお答えすることは
−0.5094	伏見政府参考人	お答え申し上げます。　まず、私も、公務員として不適切な行為があったことを深く反省
−0.5228	野上国務大臣	御指摘のございました主要農作物種子法につきましては、昭和二十七年に、戦後の食料増
−0.7074	野上国務大臣	間伐等特措法によりまして、平成二十年の法律制定後、一定以上の森林面積を有します市
−0.8432	葉梨副大臣	お答えいたします。　佐々木先生の資料の二の品目横断的経営安定対策、これが導入され
−1.0359	水田政府参考人	お答えいたします。　委員御指摘の冊子の二ページのところに「ＥＵやアメリカの現状」
−1.5408	高鳥委員長	お諮りいたします。　ただいま議決いたしました法律案に関する委員会報告書の作成につ
−1.5771	高鳥委員長	起立少数。よって、本修正案は否決されました。　次に、原案について採決いたします。
−1.9059	大串（博）委員	今回の農中さんの議論を契機に是非いい議論をしていただきたいと思いますし、間違って
−2.2223	田村（貴）委員	私は、日本共産党を代表して、本法案に反対の立場から討論を行います。　第一に、改正
−2.5166	田村（貴）委員	私は、日本共産党を代表して、畜舎等の建築及び利用の特例に関する法律案に反対の討論
−2.6210	重徳委員	立憲民主党の重徳和彦です。　今日は矢上筆頭、先輩、同僚議員の御了解をいただきまし
−3.5215	新井政府参考人	お答えいたします。　ＯＩＥ連絡協議会は、産業界及び学界における技術者又は学識経験

付録 B

ニューラル文書ベクトルを用いて, Livedoor コーパスから $T = \{$“恋愛”, “結婚”$\}$ として抽出されたテキストのスニペット. 抽出された 262 個のテキストから, ランダムに 40 個を選択して表示している. 数字は文書番号を表す.

文書番号	スニペット
48	既婚女性の話に恋愛を学ぶ独身女性ばかりの職場に勤務するサオリさん（28 歳
50	独身男性は独女より人妻と遊びたいってホント？「結婚をしたら独身男性から
86	モテる男が選ぶ女子の条件「素敵だと思う人には、もうすでに奥さんがいる」
99	人に聞いてはいけないことお盆に帰省した沙織さん（30 歳）に「実家は楽しか
116	新しい結婚の形？「事実婚」とは実際どんな制度なのか最近メディア等で「事
156	恋の駆け引きができる女、できない女「追いかけられると逃げたくなり、逃げ
181	2011 年こそ結婚したい！独女・独男の婚活事情「婚活」という言葉が流行語大
186	仕事、遠距離恋愛……。会えない時間で愛は育つか？俳優の向井理さんとモデル
239	アラフォーだって結婚式したい！ Presentedby ゆるっと cafe 独女の皆さま、はじ
298	意外に大変?!モテる定説“マメな男”との交際事情メールや電話で連絡を怠らず
309	独女的映画レビュー vol.7『』“人を好きになる気持ち”ってどんな気持ちだっけ
314	あなたはいくつまで恋ができると思いますか？今年 4 月、2 週に渡って朝日新聞
316	今さらながら、運命の出会いについて考えてみる「もしあの時……」と選ばなか
395	ケンカをしても気持ちが冷めている…。「長すぎた春」の予感先日、28 歳の誕生
448	独女通信が見た「独女たちの 5 年の軌跡」Vol.5〜婚活とは、人生最大の営業活
468	恋愛感情がイマイチでも結婚はできるのか？婚活ブームの影響で、寸前と言わ
596	1 回限りの関係で終わってしまう女昨年から婚活に励んでいる恵美さん（36 歳
618	【オトナ女子のリアルな悩み】同時に 2 人好きになってしまいましたたとえ片想
631	このままじゃ婚期を逃す!?趣味に走るオタク独女ここ数年「価値観が多様化し
650	同世代の長谷川理恵の破局について思うこと恋アラフォー、長谷川理恵（38 歳
798	結婚は“相手”より“タイミング”の問題？「」“真面目で清楚なお嬢様タイプ”の
845	独女的『「ゆるオタ君」と結婚しよう』婚活に疲れ気味、あるいは「いい男が
3435	SDN48 芹那が激白「彼氏を作って独り身を“脱出”したい」彼氏を作りたい！ SDN
3997	“運命の人”を一発検索！あなたと相性の良い生年月日はいつ？「自分の運命の
4081	内田有紀が就活＆婚活「30 ハケン女が就職する方法」エイベックス通信放送が
4178	インタビュー：眞鍋かをり「作戦を立てて一気に」現在、すっかり定着した感
4228	石田純一が恋のお悩みをサポート連載コラム「恋に効く名言」スタート 2009 年
4440	インタビュー：小原さん「自分からプロポーズすることは 100 %ない」過去に 1
4482	「グレアナ」に見る男脳と女脳、違いを理解すれば仕事も恋もうまくいく？何
4603	″ 姉妹″ の過去も!?倖田來未さん結婚今年話題の恋愛・スキャンダルニュース＜
4715	恋人には言えない本音／アソコの臭い対策法など－【恋愛】週間ランキング Pea
4722	アレが大きい彼氏とのお悩み対策法／うない元カレへのメール法など－【恋愛】
4773	男が好む清純派の演じ方／「結婚はムリ」と男に見切りをつけた瞬間など－【恋
4824	″ ゴムなし行為″ は健康にいい恋人の愛情を確かめるにはなど－【恋愛】週間ラン
5727	【SportsWatch】モデル、スピードスケート清水との交際を語る日本テレビ「の
6008	【SportsWatch】石井慧が電撃離婚か？「女性自身」が報じる 3 月 29 日発売の「
6871	稲本＆田中美保、交際順調で結婚も視野に昨年 1 月、ホテル従業員のツイッター
7013	「蒼井に似てる」大森南朋の結婚相手に突っ込み 14 日、主要メディアは俳優・
7142	林家三平の元カノ＝ NHK 荒木アナが人生を嘆くも「別れて正解」15 日、Web 版「
7212	元「モーニング娘。」高橋愛に熱愛スクープ。ファンから悲痛な叫び明日 15 日