コンピュータによる文書分類の性能評価実験

山　田　　耕

概　要

本研究は、近年著しい発展を見せているコンピュータによる自然言語処理の分類性能を実験的に確かめたものである。ジャーナリズム研究では、複数の文書をその内容に応じて少数のカテゴリーに分類して、誰が何をどのように言及しているのかを探る研究が盛んに行われている。この時の分類は、人が読んで行われることが多い。しかし、コンピュータによる最近の言語処理は、文書の内容をある程度正確に推定することができるようになってきた。それを応用すれば、内容の似た文書をコンピュータによってまとめることも可能である。本研究では、そのような言語処理の中で潜在的ディリクレ配分法とサポートベクターマシーンに注目して、それがどのくらい文書の内容に応じて分類できるかを入力データの性質という観点から考察した。その結果、文書間の関連性が弱い場合は、人が読んで分類した時と同程度の正確さで分類できることがわかった。一方で、文書間に強い関連性がある場合はまだ人の分類の方がいくぶん精度が良い。さらに、文書の主張が肯定的か否定的かを判定する分析では、人の分類の方が精度として10%程度高いことがわかった。

1　はじめに

コンピュータ、とりわけインターネット通信の発展は、テキストデータの流通量および蓄積量において爆発的な増加をもたらした。その増加と共に、大量のテキストデータを解析する手法も開発されてきている。本稿では近年注目され、目覚ましい発展を遂げているコンピュータによる自然言語処理に焦点をあてて、それらによってどの程度文書の内容を把握することができるのかを調査する。

コンピュータによる自然言語処理の目的は、簡単に言えばコンピュータが自然言語（日本語や英語など諸言語）を理解し、文の分析や生成をしてくれることである[1]。1990年以降、確率論や機械学習研究の発展の他にコンピュータの凄まじい性能向上によって自然言語処理は実用的なレベルにまで達した[2]。さらに、インターネットの登場によって電子化された文章（電子書籍、SNS 上の文章、各種のアーカイブテキストデータ）などが大量に流通し始め、容易に入手可能になった。そのため、ネット空間に蓄積された膨大なデータを効率的に、かつ正確に解析して、それらに内在している意味や知恵を抽出する必要性が高まっている。このように研究の積み重ねと社会的なニーズを背景として、近年コンピュータによる自然言語処理は非常に大きな進歩を見せている。

自然言語を理解して分析する方法として、大きく分けて４つのステップがある[2]。基本的なものとしては品詞などを決定する形態素解析があり、それをベースとして単語間の構文的関係を決める構文解析、単語や文の意味を理解する意味解析、複数の文の意味を理解する文脈解析と続く。これらの分析方法はベイズ統計を基に構築されており、それを行うコンピュータソフトも開発されてきている。例えば、日本語の形態素解析をするフリーソフトは茶筅[3]やMeCab[4]などがある。また、意味解析や文脈解析に対応するような解析は潜在的ディリクレ配分法（Latent Dirichlet Allocation、LDA と略す）[5]やサポートベクターマシーン（Support Vector Machine、SVM と略す）[6]などがあり、それらを行うプログラムも無償で配布されている[7][8]。LDA は文章に潜んでいるメッセージを推測する方法であり、SVM は判別分析の一種で文書集合を２つの集合に分類する方法である。両者とも近年急激に使われ始めている。コンピュータによる解析は、短時間で膨大なデータを処理することができ、また、結果の再現性はロバストである。このようなコンピュータによる解析のメリットをうまく利用することで、人が読んで文書を分類するヒューマンコーディングの解析と比べてはるかに大量のデータを効率的に分析することができる。

ジャーナリズム研究の一つとして、ニュースと呼ばれるジャーナリストが発

表したメッセージを解読して、その特徴を分析するものがある[9]。この分析に対して、内容分析という手法がしばしば用いられる。Krippendorff（1989）によると、「内容分析とは、データをもとにそこから（それが組み込まれた）文脈に関して再現可能でかつ妥当な推論を行うための一つの調査技法である」と説明されている[10]。また、Zelizer（2004）は「単語、フレーズ、ストーリー、画像といったものを手掛かりに文書中に現れる回数を数えたり、事前に定義されたカテゴリーに分類したりして、潜在的な意味を明確にすることである」と述べている[9]。内容分析の定義は研究者間で若干異なる[11]が、その目的は、大きく分けると複数の文書を内容別にいくつかのカテゴリーに分けてマスメディアが何に注目して報道しているのかを知ること、または文書の内容がある事象に対してどのような態度（肯定的なのか否定的なのか）で書かれているのかを知ることに分けられる[12][13][14][15]。後者の分析を特にテキスト評価分析と呼んだりする[16]。

近年、計量的な方法で文書の内容を把握しようとする研究も精力的に行われている[17][18]。それらの多くは形態素解析でキーワード単語を抽出して内容を推測する方法である。しかしながら、このようなキーワード単語の抽出に依拠した内容把握にも欠点がある。もっとも大きな欠点は、全文書もしくはある一定の大きさの標本から全体的に単語やフレーズを抽出してそれらの関係性を見るため、個々の文書がどのような話題を述べているのかというミクロな点がわからなくなってしまうことである。そのため、形態素解析で単語やフレーズを単純にカウントするだけでは、文書に含まれる重要な話題を見落とすようなことが起こる。それを回避するためには、単語間の全体的なバランスから内容を推測していく解析方法を適用しなければならない。この解析にはLDAやSVMがもっともよく使われている。本稿の目的は、文書の内容把握に対してLDAとSVMがどの程度使えるのかを定量的に検討することである。

LDAに対する過去の性能評価実験として、白井・三浦（2011）は朝日新聞、読売新聞、毎日新聞の社説からそれぞれの社説がどこの新聞社のものかを

LDA で推定し、その結果と正しい答えを比較する実験を行った[19]。それによると、60%後半の正解率で社説がどこの新聞社かを LDA は当てている。一方、SVM に対する性能評価には、高須・相原（2003）[20]がある。彼らは論文がどの学会誌・研究会誌に掲載された論文かを分類する実験を行い、80%後半の精度で分類できることを報告している。これらの研究は、主に LDA と SVM それ自体の性能に注目した実験であり、入力データは単一の条件で計算している。しかし、これらの性能は入力するデータの状態にも影響を受ける。トピック推定や二値分類などは文書固有の単語（これを特徴語という）を手掛かりに行うため、特徴語をうまく取り入れた入力データを作ることは性能を改善するのに必要である[21]。したがって、本研究では入力データの条件を色々と変えて、LDA と SVM の性能がどのように変化するかを詳しく見ていく。さらに、ヒューマンコーディングも実施し、コンピュータコーディングがどの程度人の解析に近いのかを明らかにする。このようなヒューマンコーディングとの比較報告は今までほとんどない。本研究ではマスメディア研究で使うことを想定しつつ、文書のコンピュータコーディングの可能性について有益な情報を提供したい。

2　計算方法と実験資料について

2.1　潜在的ディリクレ配分法（LDA）

一つの文単位（文章、段落、文全体など）の中には潜在的に複数の話題（トピック）が混在しており、それを基に単語が生成されていると仮定するモデルをトピックモデルと言う[5]。この仮定の下で定式化されたのが、近年注目されている LDA である。文書はあるトピック（これは文書内に隠れているので潜在トピックと言う）の下で書かれるため、文書に登場する単語の頻度はトピックに応じて変化すると考えられる。この考え方をベースにして、単語の出現頻度分布から潜在トピックを推定するものである。この LDA には様々なモデルが提案されているが、ここでは Bei ら（2003）[22]に従って概要を説明する。あ

る文単位におけるトピック割合とトピックに対する単語の出現確率分布はディリクレ分布に従い、一方で一つ一つの潜在トピックや単語は多項分布で確率が生成されると仮定する。このようにテキストデータ（今の場合、単語）の出現過程を確率変数を用いて記述したものを確率的生成モデルと呼ぶ。

今、文書数を M、文書 d の総単語数が n_d からなる文書集合を考える。ここで、文書とはある一定の意味を構成する単語の集団と定義する。全体の文書集合は複数の潜在トピック（ここでは K 個あるとする）から成り立っていると仮定され、k 番目のトピックが文書 d で出現する確率を $\theta_{d,k}$ で表す。また、同一の単語でもトピックが異なれば、その出現確率が変わる、つまり使われやすい、もしくは使われにくいといったことが起こる。例えば、「野球」という単語を考えた時に、政治ネタよりもスポーツネタで使われる確率が高くなると考えることはおかしなことではない。単語 v（全文書で V 個の異なる単語が使われているとする）が k 番目のトピックの下で出現する確率を $\phi_{k,v}$ とする。このような状況下で、確率ベクトル $\boldsymbol{\theta}_d = \{\theta_{d,k}\}$ と $\boldsymbol{\phi}_k = \{\phi_{k,v}\}$ の生成に対してディリクレ分布を仮定する。例えば、潜在トピックとして政治、経済、科学の３つを考える。この時、$\boldsymbol{\theta}_d$ は {政治、経済、科学} = {0.3,0.5,0.2} とか {0.8,0.1,0.1} といったものに対応して、この組み合わせが生じる確率が

$$\left.\begin{aligned} \boldsymbol{\theta}_d &\sim \mathrm{Dir}(\boldsymbol{\alpha}) \quad (d = 1, 2, \ldots, M) \\ \boldsymbol{\phi}_k &\sim \mathrm{Dir}(\boldsymbol{\beta}) \quad (k = 1, 2, \ldots, K) \end{aligned}\right\} \tag{1}$$

に従うというものである。(1) 式において、ディリクレ分布を Dir として表した。また、$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ は K 個の要素をもつベクトルで、$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_V)$ は V 個の要素を持つベクトルである。これらはそれぞれトピックと単語出現に関するディリクレ分布を支配するものであり、データから学習され値が与えられる[5][23]。

次に、文書 d における i 番目の単語を $w_{d,i}$ (ここで、$i = 1, \ldots, n_d$) として、その単語の背後にある潜在トピックの種類をトピック番号 $z_{d,i}$ ($\in 1, 2, \ldots, K$) で表

す。この時、$z_{d,i}$ の値が同じ単語は同じトピックに所属しているとみなされ、それらの出現分布は同じ分布に従うこととなる。変数 $w_{d,i}$ と $z_{d,i}$ は共に離散値を取るので、

$$\left.\begin{aligned} z_{d,i} &\sim \mathrm{Multi}(\boldsymbol{\theta}_d) \quad (i = 1, 2, \ldots, n_d) \\ w_{d,i} &\sim \mathrm{Multi}(\boldsymbol{\phi}_{z_d}) \quad (i = 1, 2, \ldots, n_d) \end{aligned}\right\} \tag{2}$$

という $\boldsymbol{\theta}_d$、$\boldsymbol{\phi}_{z_d}$ に依存した多項分布に従うことを仮定する。(2) 式の Multi は多項分布を表す。この一連の流れを視覚的に表したものが図 1 である。LDA の計算はフリーソフト **MA**chine **L**earning for **L**anguag**E** **T**oolkit（通称、MALLET）を使う[7]。計算の際に、トピック数を事前に決めることが要求される。また、LDA の計算では初期値に乱数を使うため同じパラメータであっても推定結果が異なる。そのため、ここでは同じ条件下で計算を三回行ってその平均値 $\bar{\theta}_{d,k}$ を用いる。

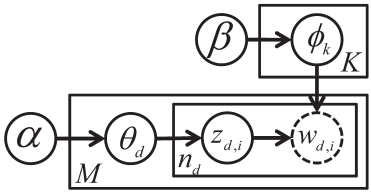

図 1：LDA のグラフィカルモデル。矢印の方向は生成の順序を示す。四角の囲みはその中にある生成過程を何回か繰り返すことを示す。点線の丸で囲まれた変数は観測データである。観測データを使って、生成の上位にある変数を推定していく。参考資料[24]を基に作成。

2.2　サポートベクターマシーン（SVM）

SVM は2000年辺りから自然言語処理において急激に使われ始めた方法で、訓練データから機械が学習してテストデータを２つのグループのどちらかに分類する二値分類器である[6]。教師あり判別分析の一種であり、訓練データから２つの集団を分ける面（分離面）を決定して、テストデータがそれぞれどちら

のクラスであるかを推定する。データ群を２つに分離する分離面は、その面とそれに最も近いデータ点までの間の距離が最大になるという条件を課すことで求められる[20]。

今、M 個の文書を考え、そのそれぞれの単語ベクトルを $\boldsymbol{x}_i$ とする。また、文書一つ一つは肯定的か、否定的かのどちらかに分類されていると仮定する。この時、分離面を表す関数を次のように定義する：

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{3}$$

ここで、$\boldsymbol{w}$ は分離面の法線ベクトルで、b は切片である。一般性を失うことなしに、

$$\min\{|f(\boldsymbol{x}_i)|, i = 1, \ldots, M\} = 1 \tag{4}$$

という条件を $\boldsymbol{w}$ と b に課す[25]。この時、分離面に最も近い点までの距離は $1/|\boldsymbol{w}|$ となる。SVM では、この距離が最大になるように $|\boldsymbol{w}|$ を決めることで分離面が求められる。態度不明の文書の単語ベクトル $\boldsymbol{y}$ を考えた時に、$f(\boldsymbol{y}) \geq 0$ ならば肯定的、$f(\boldsymbol{y}) < 0$ ならば否定的と判定される。ここで、分離面は入力する文書の単語ベクトル $\boldsymbol{x}_i$ によって変わることに注意が必要である。

本研究では、フリーソフトの SVM-Light[8] を使う。このソフトは、成分が何千にもなるベクトルを扱うことができ、数万の訓練／テストデータにも対応している。このアルゴリズムには、単語ベクトルを生成する際によく見られる高次元のスパースなベクトル、すなわち、０の成分を多く持つベクトルのデータに対しても最適に計算してくれる方法が採用されている。また、一般的に分離面は超平面ではなく、複雑に曲がりくねった超曲面である可能性が高いため、(3) 式をそのまま適用しても精度の良い分類は期待できない。ここでは、非線型な分類を可能にするために動径基底関数カーネルを導入する[26]。カーネル法は非線型的な分類だけでなく、計算量の増加を和らげる機能もある[2]。

2.3　実験資料について

上述したLDAとSVMの性能評価で用いる資料について説明する。LDAによるトピック推定の性能評価は朝日新聞の記事を用いて、それらの記事内容をどの程度正しく推定できるかで測る。記事抽出は朝日新聞「聞蔵Ⅱ」に以下の各検索キーワードを入力して抽出する。ただし、抽出する面に対する制限は付けない。様々な用途で使うことを想定して、ここでは検索キーワードに引っかかったものを解析対象として扱う。ただし、カテゴリー間で重複する記事は除外した。

実験は異なる資料を使って２つ行った。実験１では一つ一つの記事の間にあまり関係がないものを選んだ。すなわち、一つの記事内で他方が同時に言及されることがあまりないという基準で記事を収集する。これはもっとも分類がしやすい場合に対応する。一方、実験２では、あるキーワードの下で語られている記事をどの程度分類できるかを見たものであり、推定が難しい場合に対応している。両ケースでの性能を評価することで、LDAの推定精度範囲を測定することができる。

最初に、実験１では2015年４月10日を起点として過去にさかのぼる形でそれぞれ150件の計750件の記事を選んだ（すべてアンド検索）：

1. ウクライナ＆ロシア＆米国【ウクライナ】
2. テニス＆錦織＆大会【テニス】
3. 原子力＆安全＆審査【原子力】
4. 集団＆政府＆自衛権【集団】
5. 小保方＆理研＆研究【STAP】

【 】内の単語は検索キーワードで抽出された記事のトピックを代表するトピックワードと定義する。実験１の記事の平均文字数、標準偏差、最小文字数、最大文字数は表１の通りである。表１の平均文字数を見るとどのトピックも大

体1000字程度であるが、テニストピックに属する記事がやや短い。標準偏差を見ると1000字程度となっており、記事の大部分は100字から2000字以内のものであることを示唆している。文字数の短い記事は抽出される単語数が少ないため、一方、長い記事はその中に様々な内容（トピック）が含まれてくるため、両方とも推定精度が悪くなる可能性がある。今の場合、そのような300字以下の短い記事は68件、4000字を超す非常に長い記事は15件である。

表１：実験１で使用する記事の各トピック所属記事の統計量。文字数はデータベースに記載されている数字を使った。

トピック	記事数	平均文字数	標準偏差	最小文字数	最大文字数
ウクライナ	150	1468	1335	186	7853
テニス	150	997	826	94	4890
原子力	150	1136	956	176	7079
集団	150	1437	1241	227	11423
STAP	150	1037	783	104	3935

次に、実験２では2015年８月19日を起点として過去にさかのぼる形でそれぞれ150件の計450件の記事を選んだ（すべてアンド検索）：

1. 地震 & 原発 & 事故【原発】
2. 地震 & 津波 & 避難【津波】
3. 地震 & 防災 & 対策【防災】

実験２では、地震というキーワードの下で記事が選定されている。これらの記事の平均文字数、標準偏差、最小文字数、最大文字数は表２の通りである。防災トピックの記事は他の２つのトピックに比べて平均文字数がやや多い。ただし、標準偏差は1000字程度なので記事の大部分は100字から2000字以内のものであり、実験１の記事と同じような文字数分布を示している。また、300字以下の記事は13件、4000字以上の記事は４件である。

表2：実験２で使用する記事の各トピック所属記事の記述統計量。文字数はデータベースに記載されている数字を使った。

トピック	記事数	平均文字数	標準偏差	最小文字数	最大文字数
原発	150	1214	1008	227	9549
津波	150	1160	628	166	2879
防災	150	1617	1135	117	10589

次に、SVM に対して用いる資料について説明する。SVM はある種の評価をすでに含んだ資料が性能評価実験には適している。そのため、ここでは書籍に関する評価がなされている AMAZON のカスタマーレビューを使う。評価は５段階で５はもっともおすすめできる本であることを示しており、１はその逆である。本研究では、以下で挙げた書籍（タイトルと著者、評価５と評価１の2015年６月27日時点のコメント数）のカスタマーレビューを使う：

- 21世紀の資本、トマ・ピケティ（評価値５は65件，評価値１は８件）
- 捏造の科学者 STAP 細胞事件、須田桃子（評価値５は43件，評価値１は13件）
- イスラーム国の衝撃、池内恵（評価値５は41件，評価値１は３件）
- 統計学が最強の学問である、西内啓（評価値５は49件，評価値１は25件）
- 生物と無生物のあいだ、福岡伸一（評価値５は171件，評価値１は24件）
- 学年ビリのギャルが１年で偏差値を40上げて慶應大学に現役合格した話、坪田信貴（評価値５は149件，評価値１は82件）
- 家族という病、下重暁子（評価値５は17件，評価値１は64件）
- 「少年Ａ」この子を生んで……―父と母悔恨の手記、「少年Ａ」の父母（評価値５は19件，評価値１は48件）

様々なジャンルの本に対するカスタマーレビューを使用することで、分類性能が特定の話題に依存する可能性を排除することができる。これらのカスタマー

レビューで評価５のものは554件、評価１のものは267件である。評価５と１のレビューをそれぞれ文字数の多い順に200件抽出する。その後、ランダムサンプリングで２つに分けて片方を訓練データとして使い、もう片方をテストデータとして使う。訓練データ（200件）とテストデータ（200件）それぞれの平均文字数および標準偏差は569±586字と642±926字となる[1]。若干テストデータの方が平均文字数が大きいが、有意水準５％で両側検定の t 検定を行ったところ $t(398) = 0.936$ で有意確率0.35となり、両者の平均文字数の差に有意差は見られなかった。カスタマーレビューの高い評価である５の記事は肯定的な例として、評価の低い１の記事を否定的な例として扱う。訓練データからテストデータのテキスト評価を試みる。

LDA および SVM では、各文書に対して出現単語リストと単語ベクトルを入力データとして使うため文書の形態素解析を事前に行わなければならない。そのため、形態素解析には KHcoder[27] を使って単語を抽出し、Okapi BM25の方法で重み付けを与える[28]。Okapi BM25は、一部の文章にたくさん出てくる単語に対して数値が高くなるように重み付けを与える指標生成法であり、出現数だけでなく、全文書のうちどのくらいの文書に登場しているのか、その単語がそれを含む文書内に占める割合などを勘案して算出される。つまり、Okapi BM25では「する」や「ある」といったどの記事にも出てくるような単語（一般語と呼ぶ）に対し非常に小さい値に変換され、一方、特徴語については大きな値となる。これにより、特徴語を容易に見つけ出すことができる。本研究で Okapi BM25を導出する際に、文字数はその文書に含まれている抽出単語数で代用し、その他のパラメータは吉岡・小枝（2012）[21]にならった。

本研究では、上述したコンピュータの性能がヒューマンコーディングと比較した時に相対的にどの程度であるのかも調べる。そこで、早稲田大学政治経済学部所属の学生３人にコーダーとなって分類を行ってもらった。LDA のトピ

1 この文字数は KHcoder で計算される length_c を使った。詳しくは、樋口（2014）[29] を参照せよ。

ック推定との比較では、実験１資料記事750件から200件を選んだ資料を渡した。この資料は各トピックに所属する記事150件からランダムサンプリングした40件から構成されている。同じように、実験２も各トピックからランダムサンプリングした40件から構成される120件の記事を渡した。それぞれの記事を５つのトピック（テニス、STAP、ウクライナ、集団、原子力）および３つのトピック（原発、津波、防災）に分類してもらった。二値分類も同様に、評価５と評価１それぞれ50件をテストデータ（200件）の中からランダムに抽出して、100件のデータを作った。コーダーには、事前に分類方法を説明した上で作業してもらった。

３　各方法の性能評価について

3.1　LDA による性能評価結果：実験１

表１で示した記事一つ一つを一つの分析単位として、750件の記事に対して LDA によるトピック推定を行う。この時、分析にかける各記事の単語は次のように決める。各記事内で使われている各単語の Okapi BM25値が、全体の平均値の g 倍よりも大きい単語をその記事での特徴語とみなして分析にかける。パラメータ g は特徴語を決めるパラメータである。ここでは、g = 0.25の場合を標準ケースと呼ぶ。標準ケース以外にも、g = 1.25、1、0.75、0.5、0 の場合について同様の計算を行う。これらはそれぞれ一つの記事辺り平均46（g = 1.25）、74（g = 1）、110（g = 0.75）、148（g = 0.5）、171（g = 0.25）、176（g = 0）単語を使っていることに相当する。トピック数は５、計算の反復回数は1000回に設定する。

最初に、標準ケースの計算結果について述べる。LDA では、各記事に対して各トピックの確率を導出する。つまり、セクション２で述べた $\bar{\theta}_d$ に対応する。それと同時に各トピックに対する代表的な単語も算出される（$\phi_{k,v}$ の大きい単語）。標準ケースでは、表３のようになる。トピック数が５であるため、トピック番号は０から４までとなり、その各々に代表的な単語が出力されている。

各トピック番号の意味付けは、この代表的な単語群を手掛かりに人が判断しなければならない。今回の場合、各トピック番号に対して表３のようにトピックワードと潜在トピックを紐付けた。

表３：標準ケースにおけるトピック番号にリンクされた代表的な単語群。

トピック番号	単　　語	意味付けたトピックワード
0	原発、稼働、月、規制、審査、安全、事故、原子力、同意、対策、委員、必要、計画、判断、国、求める、地元、基準、福島	原子力
1	錦織、大会、世界、選手、決勝、優勝、日本、テニス、男子、圭、試合、ランキング、シード、シングルス、全米、オープン、相手、自分、東京	テニス
2	ロシア、ウクライナ、米国、日本、経済、大統領、中国、政権、プーチン、関係、国際、制裁、合意、eu、会議、欧州、米、世界、停戦	ウクライナ
3	細胞、研究、小保、論文、stap、理研、月、実験、調査、科学、いう、できる、委員、問題、検証、発表、不正、センター、教授	STAP
4	自衛隊、政府、日本、言う、米、安倍、自衛、集団、支援、首相、反対、事態、行使、協議、月、活動、決定、保障、与党	集団

推定の精度を見る前に、各記事に付加される確率 $\overline{\theta}_{dk}$ について以下の記事Ａを例として説明する。

記事Ａ：STAP 特許の出願継続　理研　【大阪】

理化学研究所は、国際出願していた STAP 細胞作製法の特許について、複数の国で出願の継続手続きをとった、と24日明らかにした。論文は撤回されたが、STAP 細胞の存在は完全に否定されてはいないとし、検証実験も続けていることから、特許取得の手続きを進めることにしたという。国際出願は条約加盟国すべてに特許の出願をした効果があるが、実際に特

許を得るには各国でそれぞれ手続きをとる必要がある。24日は、国際出願から各国への手続きに移行できる期限だった。理研は、移行手続きをした国を明らかにしていない。今後、特許出願した内容が、各国で審査される。発明者には、理研発生・再生科学総合研究センターの小保方晴子氏も含まれている。

（朝日新聞、2014年10月25日、朝刊、５総合、308字）

この記事の $\overline{\theta}_{\mathrm{A},k}$ は表４のようになる。今、確率の高い順に数えた時の i 番目の確率を P_i とする。記事Ａでは、STAP トピックを表す $\overline{\theta}_{\mathrm{A},3}$ が一番高くなっているので、$P_1 = \overline{\theta}_{\mathrm{A},3}$ となる。

表４：記事Ａに付加された確率。確率の高い順に並び替えている。

トピック番号	確率 $\overline{\theta}_{\mathrm{A},k}$
3	0.502
2	0.188
0	0.151
4	0.089
1	0.069

次に、全記事に対する推測の精度について検討していく。本研究では、各記事の主要な内容は一番高い確率 P_1 が与えられたトピックであると見なす。すなわち、$P_1 = \max\{\overline{\theta}_{d,k}, k = 0, 1, 2, 3, 4\}$ となったトピック番号 k をその記事の主要な内容とする。記事Ａの例では、STAP トピックの確率が一番高かったのでその記事の内容は STAP に関することが書かれていると考え、小保方＆理研＆研究のキーワードから抽出された記事であると判断する。このようにして LDA が推定した各記事のトピックが得られ、これと実際の正しいトピックを対比することで性能評価をすることができる。標準ケースでは、750件の記事のうち722件（96.3％）で元の正しいトピックと一致した。各トピックに対す

る内訳は表５のようになった。表５は表頭に実際のトピックを、表側にLDAで推定されたトピックが配置されている。精度は、LDAでそのトピックであると推測された記事のうち実際に正しいトピックに所属する記事数との比で定義されている。テニストピックを見ると、LDAでは155件がテニストピック所属記事と推測されているが、実際に正しいテニストピック所属記事は149件である。そのため、精度は149/155 = 0.961となる。表５から、ウクライナトピックの精度は他のトピックに比べて悪くなっている。また、集団トピックは精度は高いが、推定が成功した記事は他のトピックよりも少ない。集団トピック所属の記事のうち、いくつかはウクライナトピックの内容に似たもの、つまり国際関係の中で集団的自衛権が語られるという記事があったためである。

表５：標準ケースにおける正しいトピックとLDAで推測されたトピックのクロス表および各トピックに対する精度。

		正しいトピック					精度
		ウクライナ	テニス	原子力	集団	STAP	
LDAで推測されたトピック	ウクライナ	141	1	0	10	2	0.916
	テニス	4	149	0	1	1	0.961
	原子力	1	0	148	1	0	0.987
	集団	3	0	1	137	0	0.972
	STAP	1	0	1	1	147	0.980

では、どのような記事において推定の失敗が起こるのであろうか、その原因について標準ケースを例に考えていく。表６は各記事の $X = (P_1 - P_2) / P_1$ を計算して、分類したものである。X が１に近づくということは $P_1 \gg P_2$ であることを意味している。表６の第２列目に推定に成功した記事数 N_s が、第３列目に推定に失敗した記事数 N_f が表示されている。第４列目には、推定に失敗した記事の割合 $N_f / (N_s + N_f)$ がパーセント表示で示されている。この割合の変化から、X が大きくなると急激に推定に失敗した記事の割合が減少していることがわかる。具体的には、X が0.6までは指数的に減少している一方で、$X > 0.6$

でほぼ一定になっている。この急減は推定に成功した記事が X の増大と共に爆発的に増加していることに起因している。表６から、推定に失敗する割合は X が小さいところで高く、X が大きくなるにつれて低くなると結論付けることができる。

表６：標準ケースにおける $(P_1 - P_2) / P_1$ に対する推定に成功、または失敗した記事数とその割合。X は0.2刻みでまとめている。

番号	範囲 X^a	推定に成功した記事数 N_s	推定に失敗した記事数 N_f	割合 (%)^b
1	$0 \leq X < 0.2$	12	6	33
2	$0.2 \leq X < 0.4$	25	9	26
3	$0.4 \leq X < 0.6$	49	7	13
4	$0.6 \leq X < 0.8$	211	2	0.9
5	$0.8 \leq X < 1.0$	425	4	0.9

[a] $X = (P_1 - P_2) / P_1$
[b] $N_f / (N_s + N_f) \times 100$

ここで、推定失敗した $X < 0.6$ の記事についてその記事の P_2 に対応するトピックが正しいトピックと一致しているかどうかを調べる。表６から、$X < 0.6$ では22件において推定失敗をしていることがわかる。その中で、P_2 が一致している記事は22件中19件（86%）になる。一方、推定失敗した $X \geq 0.6$ の記事は P_2 と対比しても、６件中４件（67%）しか合っていない。このことは、推定失敗した $X < 0.6$ の記事は２番目に確率が高いトピックが正しいものを多く含んでいることを意味している。また、文字数と推定精度の関係を見ると、300字以下の記事では68件中１件しか推定の失敗はなかった。4000字以上の記事では15件中４件において推定の失敗が見られた。長い文章の記事はその中に様々な要素が盛り込まれているため、推定失敗の確率が高くなったと考えられる。

推定精度は分析に使用する単語、つまりパラメータ g に依存する。図２は、g に対する全記事（750件）と各トピック所属記事（各150件）に対する推定の

精度を示したものである。図２(a)と(b)を見てわかるように、g に対する精度は同じような振舞いを示し、g が大きくなると推定精度が低下する。これは解析に使用できる単語数が減るためである。一方、g が小さいところでは精度の減少は見られないが、各トピックを代表する語に一般語が入り込んでくる。そのため、表３のような単語のセットが出力された時に解釈しにくくなる可能性

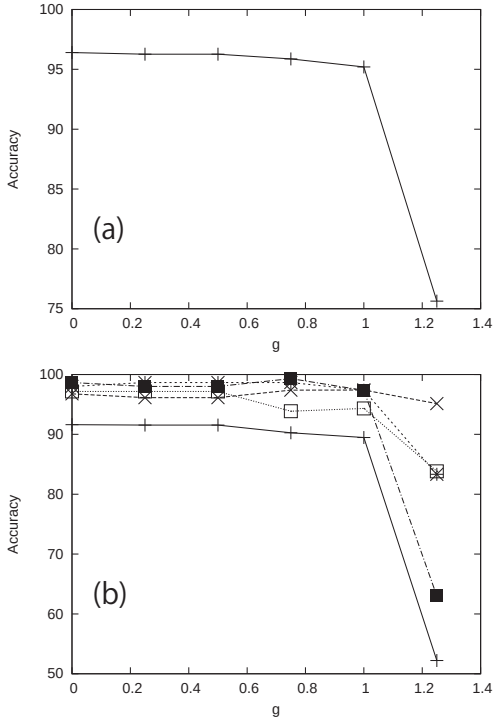

図２：g に対する実験１資料の推定精度。(a)750件中の精度(%)。(b)各トピック所属記事150件に対する精度(%)。(b)において、＋印の実線はウクライナトピック、×印を持つ破線はテニストピック、＊印を持つ破線は原子力トピック、□印を持つ点線は集団トピック、■印をもつ一点鎖点は STAP トピックに対応している。ウクライナトピック以外はほとんど線が一致している。(a)と(b)からわかるように、$g = 1$ まで精度は飽和している。g が大きくなると解析に使用される単語数が減るため、その精度が落ちる。$g = 1.25$ は一つの記事あたり46単語を使って記事の内容を推定している。

がある。例えば、$g=0$ の時、STAP トピックを表しているだろう単語群は「する、細胞、研究、小保、論文、stap、理研、月、ある、実験、調査、科学、いう、できる、委員、なる、問題、検証、発表」となり、表3と比べて「する」、「ある」、「なる」などの一般語が入ってくる。この実験から、LDA にかける際に一つの記事あたり平均100単語を下回ると精度が悪くなると言える。

実験1の資料に対するヒューマンコーディングの精度（95%信頼区間）は96.7±7.1%である。図2と比較すると、$g<1$ のコンピュータコーディングの結果はヒューマンコーディングの値とほとんど同じである。また、ヒューマンコーディングのトピックごとの精度を見ると、ウクライナトピックのいくつかが集団トピックと判定されており、集団トピックの精度が他のトピックに比べて低くなっている。このような推定に失敗する傾向はコンピュータと同じようである。実験1では記事間に出てくる特徴語をはっきりと区別でき、このような場合では両者の解析に大きな差は出にくいと言える。

3.2 LDA による性能評価結果：実験2

実験2の資料は、実験1の資料と比較してその内容において相互に強い関係がある。このような場合のトピック推定の精度について見ていく。解析は §3.1 と同じパラメータを使ったが、トピック数は3に再設定した。抽出する単語数はここでは $g=1$、0.75、0.5、0.25、0 とし、それぞれは一記事あたり65、97、131、155、159 単語を含む。最初に、標準ケース（$g=0.25$）の場合を考える。表7は標準ケースの場合の精度を示している。全体の精度は450件中326件（72.4%）であり、実験1と比べればその精度は落ちている。また、表7からわかるように個々のトピックに対する精度も落ちている。

相互に関連性のある文書では、どのトピックも同じような確率になると予想される。表8は標準ケースの場合の X 分布を示しており、$X<0.6$ の記事が全体の約70%を占めている。この傾向は実験1と異なるが、推定失敗率は X が小さくなると高くなるという傾向は同じである。ただし、その絶対値は表8の

表7：標準ケースにおける正しいトピックとLDAで推測されたトピックのクロス表および各トピックに対する精度。

		正しいトピック			精度
		原発	津波	防災	
LDAで推測されたトピック	原発	107	5	18	0.823
	津波	34	115	28	0.700
	防災	9	30	104	0.727

表8：標準ケースにおける $(P_1 - P_2) / P_1$ に対する推定に成功、または失敗した記事数。

番号	範囲 X^a	推定に成功した記事数 N_s	推定に失敗した記事数 N_f	割合 (%)^b
1	$0 \le X < 0.2$	45	43	49
2	$0.2 \le X < 0.4$	68	33	33
3	$0.4 \le X < 0.6$	89	33	27
4	$0.6 \le X < 0.8$	114	15	12
5	$0.8 \le X < 1.0$	10	0	0

[a] $X = (P_1 - P_2) / P_1$
[b] $N_f / (N_s + N_f) \times 100$

方が高い。次に、２番目に確率の高いトピックを見てみると、P_1でトピック推定に失敗している記事（124件）のうちP_2で正しいものは全部で97件に達する。そのうち、$X < 0.6$では109件中87件が２番目のトピックで推定に成功している。今回はトピックが３つしかないことからこのような高い一致をはじき出している可能性は排除できないことに注意が必要であるが、そのようなことを考慮しても、これは２番目のトピックが非常に重要であることを意味している。300字以下の記事と4000字以上の記事での推定成功の精度はそれぞれ13件中８件（62%）、４件中２件（50%）である。実験１の場合と違い、実験２では短い文章の記事の精度も全体と比べると低くなっている。

図３は全記事（450件）および各トピック所属記事（各150件）に対する推定精度のg依存性を示す。図２と同様に小さいgではその精度はほとんど変わらないが、gが大きくなると精度は57%程度まで落ちる（図３(a)）。図３(b)を

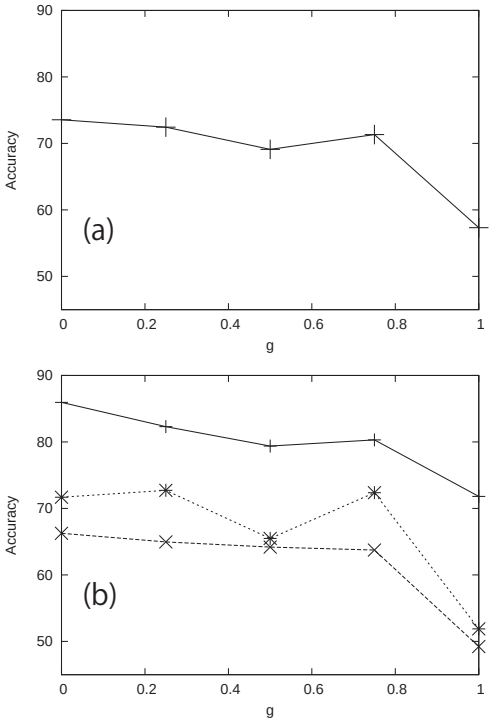

図３：g に対する実験２資料の推定精度。(a)450件中の精度(%)。(b)各トピック所属記事150件に対する精度(%)。(b) において、＋印の実線は原発トピック、×印を持つ破線は津波トピック、＊印を持つ点線は防災トピックに対応している。(a) からわかるように、$g \leq 0.75$ まで精度は大体一定である。

見るとトピックによってその精度は異なるが、g に対する傾向は図３(a)と同じようなものである。実験２の資料に対するヒューマンコーディングの精度（95%信頼区間）は83.1±17.4%となる。平均値を比較すると人が読んで分類した方が10%程度良い成績を示す。記事間に相関の高い記事分類ではまだ人の判断の方が正しくできることを示唆している。ただし、エラーバーの範囲を考慮すると、コンピュータコーディングが出した値（72%）も含まれており、両者の精度に大きなかい離があるとは言えない。各トピックのヒューマンコーディングの精度を見ると、原発トピックは88.2±6.3%（95%信頼区間）、津波トピ

ックは82.6±4.3%、防災トピックは80.3±36.9%となっている。原発トピックのヒューマンコーディングとコンピュータコーディング（図３(b)）の精度は他の２つに比べて高い。これは原発トピックが地震というキーワードの下でも他の２つのトピックとは明確に異なる内容であったためと考えられる。一方で、津波と防災トピックは様々な内容を含む分推定が難しかったと言える。

3.3　SVM による性能評価結果

最初に、訓練データとテストデータに対して、品詞分解を行って単語ベクトル $\boldsymbol{x}$ を生成する必要がある。ここでは、単語ベクトルの各成分の値に対して頻出数と Okapi BM25の数値の２通りを用意した。この際、抽出した品詞は名詞、動詞、形容詞、副詞、形容動詞、感動詞の4928単語である。精度評価計算は、４パターンの計算方法、SVM-Okapi BM25、SVM- 頻出数、判別分析 -Okapi BM25、判別分析 - 頻出数に基づいてその精度を検証した。判別分析は線型判別関数を採用して、IBM SPSS（version 20）で計算した[30]。

性能評価は次のような手順で進めた。訓練データの件数 N_t を10件から始めて10件ずつ増やしていった時の200件の各テストデータに対する評価（肯定的／否定的）の推定値と実際の評価を比較する。ただし、訓練データをもとに計算される分類規則、つまり分離面はその訓練データの単語ベクトルに依存してしまうため、精度も訓練データの組み合わせによって変化する。そのため、ここでは $N_t \leq 190$件の計算では、200件の訓練データからランダムサンプリングによって N_t 件抽出して精度を求める作業を５回繰り返した。

図４は精度計算の結果である。SVM の手法を使った場合、訓練データを増やしていくと精度が上がり、最終的に約85%に達する。また、頻出数を使うよりも Okapi BM25を使った方が、若干精度が良いことがわかる。Okapi BM25で作ったベクトルは特徴語により大きな値を与えるため、ベクトル空間内の特徴語に関する次元でより鮮明になったと考えられる。しかし、頻出数でもほぼ変わらない精度を出すことから、一般語に関するベクトル成分は今回の評価判

定においてあまり寄与していない可能性がある。一般語の次元は両者に共通して一定程度あることが寄与しない原因であろう。さらに、訓練データの件数が $N_t \geq 100$ ではほとんど一定の精度とエラーバーの大きさで推移することが図４から観測される。この精度の飽和現象は、高須・相原（2003）[20]とも調和的な結果となっている。一方、線型判別関数を使った判別分析では訓練データ数 N_t に関係なく、平均60%（頻出数）、55%（Okapi BM25）である。今の場合、テストデータには評価５と１の文書が50%ずつ入っており、すべてのテストデ

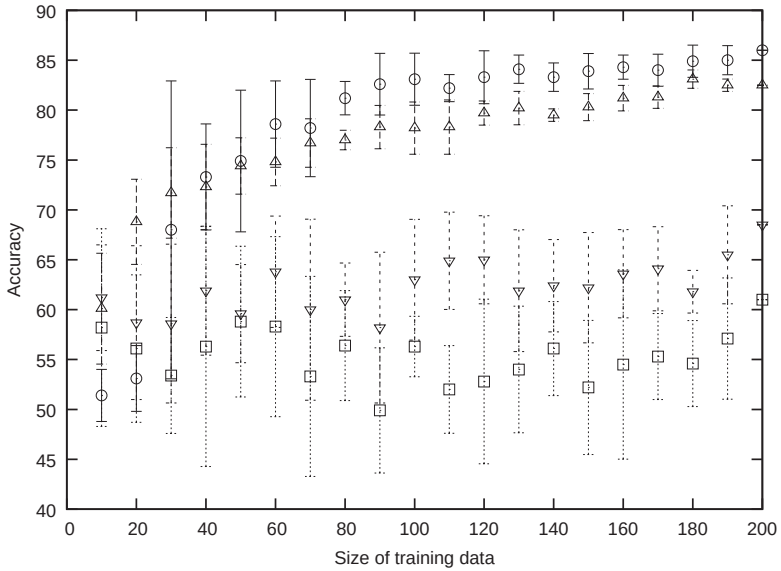

図４：横軸は訓練データの数、縦軸は精度（%）を示す。200件の場合を除くすべての場合で訓練データの標本を５回変えて精度の平均値とエラーバーを算出した。○印は SVM-Okapi BM25、△印は SVM- 頻出数、□印は判別分析 -Okapi BM25、▽印は判別分析 - 頻出数を示す。エラーバーは95%信頼区間である。エラーバーに対して○印に付いている実線は SVM-Okapi BM25、△印に付いている長破線は SVM- 頻出数、□印に付いている点線は判別分析 -Okapi BM25、▽印に付いている短破線は判別分析 - 頻出数のものである。一番精度が高いのは SVM-Okapi BM25を使った場合である。一方で、エラーバーの幅を考慮すると、訓練データが少ないところ（$N_t < 40$）ではすべての場合で同じような精度を有しているとみなせる。

ータを肯定的もしくは否定的と判断しても精度が50％になる。そのため、エラーバーを考慮すると、判別分析 -Okapi BM25はほとんど評価分析には使えないということが言える。判別分析 - 頻出数の方法は多少50％よりかは大きい値を示すが、こちらも実用に耐えられる精度に達しているとは言えない。

最後に、ヒューマンコーディングの結果との比較について述べる。ヒューマンコーディングによる精度（95％信頼区間）は95.3±3.8％になった。エラーバーを考慮しても、図４の結果と比べると人が読んで分類した精度は10％近く高いことがわかる。ヒューマンコーディングでは、文書の極性を決める特徴語を的確に掴むことができるため、このような高い精度で分類ができたと考えられる。

結論および議論

本研究では、近年注目を集めている２つの言語処理に対する入力データの影響を主に検証した。コンピュータによる言語処理を使えば、従来ヒューマンコーディングではできないような大量の文書を分類することができる。その上、解析の大部分は数学的なルールに則って行われるようになるため、追試が簡単にできるようになり、結果の信頼性も高まる。本研究では、入力データの条件を色々と変えて、LDA および SVM の推定精度がどのように変わるのかを定量的に調べた。

新聞記事を対象とした LDA の推定ではその内容を文書間に強いつながりがないデータ集合は約96％の精度（$g \leq 0.75$）で正しく推定することができ、文書間に強いつながりがあるデータ集合は70％程度（$g \leq 0.75$）で正しく推定できることがわかった。しかし、実験２の平均精度はヒューマンコーディングに比べれば10％程度劣るものであった。一般的に、精度は特徴語をどのように抽出するのかという基準、ここでは g に依存しているが、実験１と２共に、g が小さい値では全記事だけでなく各トピックに対してもその正解率は g の値に関係なくほぼ一定値となる。ただし、単語をすべて使うと一般語が解析に多く含

まれることとなる。これらは時としてノイズとなり、トピックを表す単語群の解釈が難しくなる可能性がある。さらに、多くの単語を取り入れると計算コストがかかるというデメリットもある。一方で、g を大きくして特徴語を絞りすぎると精度が落ちる。本研究から最低でも一記事あたり平均100単語以上（重複を含む）ないとそれなりの精度が出せないことがわかった。この単語数ぐらいないと文書内の単語出現確率とトピック確率を精度良く関係付けられないのであろう。

本研究で用いた資料に対して、今回２つのケースを想定して記事を抽出した。図５は、文書間の関係性を示した概念図である。実験１は図５(a)にあたり、実験２は図５(b)に対応する。白井・三浦 (2011) の実験[19]も図５(b)にあたる。一般的な内容分析ではある一つの事象、例えば、STAP 問題やウクライナ問題の文書に関心があった場合、関連検索キーワードの下で抽出した文書集合を分類することを目的に行われる。つまり、図５(b)のような文書集合を解析することが想定される。このような集合では、似たような特徴語が多くの文書で出現する可能性がある。この時、LDA の計算で言えばトピック間の確率に差が出にくくなると予想される。実際、実験２では X の小さい記事が多くなっていることが確認された。それに伴ってトピック推定の精度も落ちている。また、記事分類に対する精度の安定性、すなわち図５(a)から(b)までの精度の変化幅はヒューマンコーディングの方が小さい（96.7% → 83.1%）。図５(a)の場合ではどちらもほぼ同じ精度を出すが、図５(b)では LDA の精度はヒューマンコーディングよりもある程度劣る可能性は否定できない。しかし、このことは必ずしも常にヒューマンコーディングの精度が上であるということを意味している訳でもない。実際、今回の解析では LDA と同じような精度を示すヒューマンコーディングもあった。ここで、重要なことは誰が読んでもそのトピックに分類するという文書をコンピュータも間違わないで分類することであろう。この点については、実験１の結果および実験２の原発トピックの推定精度がヒューマンコーディングのそれとあまり変わらなかったことからある程度担保さ

れたのではないかと考えている。一方で、微妙な、つまり、読み方によっては落とし込むトピック番号が変わる文書をどう正しく見分けるかは悩ましい問題である。現実的には、文書の正しいトピック分布は（誰にも）わからないし、ヒューマンコーディングの結果が常に正しいとも限らない。しかしながら、§3で述べたようにLDAの推定失敗はXの小さい所で多い（表8）。このことはそのような文書に対するLDAの分類はまだ信頼性が低く、より良い分類分布を得るためにはヒューマンコーディングでの再チェックがXの小さい文書では必要であることを意味しているだろう。

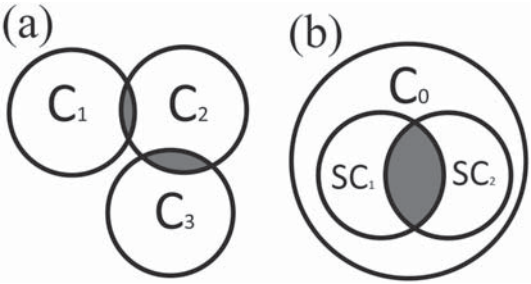

図5：文書間の関係性の概念図。(a)文書間にあまり関連性がないため、特徴語の重複はほとんどない。(b)あるクラスター(C_0)の中にあるサブカテゴリー（ここではSC_1とSC_2）間では、特徴語の重複がそれなりにある。

さらに、トピック推定に対する初期値の影響について述べておく。本研究では、同じパラメータの計算を3回行って、各文書に付される$\theta_{d,k}$の平均値をとった。この値を使った推定精度は、各回の結果と比べると良いことが確認された。これは平均値を使うことで、たまたま失敗してしまったというケースを減らせたためと考えられる。したがって、LDAの計算は1回だけではなくなるべく複数回繰り返した後の平均値を使うべきである。また、gに対して精度が一定になっている範囲において各文書の$\theta_{d,k}$をgで積分した値を使っても、本稿で示した結果と同程度の精度を出していることが観察された。このことから、gで積分した$\theta_{d,k}$を用いてもそれなりの推定ができる。

二値分類は、Okapi BM25を使ったSVMによって85％程度の精度で肯定的な文書と否定的な文章を推定することができた。この際、SVMで十分な精度を出すためには訓練データは100件程度必要であるが、それ以上の訓練データを用意しても精度向上はあまり期待できないこと、また、単語の頻出数を使うよりかはOkapi BM25を使った方がいくぶん正確な二値分類ができることを明らかにした。しかし、SVMによる分類においても文書の性質は大きく推定精度に影響を与えていることは否めない。本研究で使用した資料は、アマゾンのカスタマレビューのうち評価が１と５のものである。これらは評価の中では両極端に位置するものであり、図５(a)のように文書間に強い関連性がなく、内容に大きな隔たりがあったと考えられる。図５(b)のような文書集合でどのような精度を出せるかどうかチェックすることは今後の課題である。精度を向上させるキーポイントはどのような単語群を解析に用いるかにかかっている。文書の極性に大きな影響を与えているのは単語の感情極性である。単語の感情極性とは、「素晴らしい」、「悪い」といったポジティブ、ネガティブな意味を持っているものであり、主に形容詞がそれにあたる[31]。本研究では、一つ一つの文書の文字数が600字程度であり、形容詞のみを使うと単語ベクトルがかなりスパースになるため、形容詞以外の品詞も使った。もちろん名詞や動詞などが一概に悪いということはないが、感情極性をあまり持たない品詞を多く入れることは文書の極性が見えにくくなるという欠点があることは注意しなければならない。最近では、日本語評価極性辞書といったデータも整備されてきている[32][33]。このようなデータを使いながら、感情極性を示す単語を文書から多く抽出することが精度向上には必須である[34]。

本稿で見てきたようにコンピュータの助けを借りて分析を行うことは、結果の再現性および信頼性、大量にデータを短時間で、かつ一貫したルールの下で分析できる点でヒューマンコーディングを圧倒している。もちろん、比喩などを含む文学的な文書の理解においてまだコンピュータの解析はヒューマンコーディングに劣っている。そのため、そのような部分は人が読んで補完するしか

ない。しかし、現状でも単語の重み付けをかますことでヒューマンコーディングと遜色ない値を出すことができる場合がある。本稿では入力データの条件に着目して精度に対するその影響を論じてきた。入力データをうまく工夫することで推定精度のより高い解析ができるようになる。過去数十年間におよぶ言説の変化、Web を含む様々なメディア媒体上での言説の比較など大量のデータ分析を効率的に行うことは、メディア研究におけるメタ研究への道を大きく開く上で非常に重要となるだろう。

謝辞

本研究にあたり、目黒光司氏（東京工業大学精密工学研究所）からはたくさんの非常に有益なコメントをいただきました。数値計算に関しては、納田明達氏（東京工業大学地球生命研究所）から有意義なサポートを多くいただきました。稲葉知士氏（早稲田大学国際学術院）からはヒューマンコーディングの方法について示唆に富む助言をいただきました。また、早稲田大学政治経済学部の山田祐基氏、鹿島雄貴氏、濱正太郎氏には本研究のヒューマンコーディングに参加いただき、内容を推定してもらいました。最後に、編集委員会の適確な指摘によって本原稿は飛躍的に改善されました。この場を借りて皆様に深く御礼申し上げます。

参考文献

[1] 田中穂積（監修）．自然言語処理―基礎と応用―，電子情報通信学会，1999.
[2] 奥村学．自然言語処理の基礎，コロナ社，2010.
[3] http://chasen.naist.jp/hiki/ChaSen/（オンライン）（閲覧日：2015年11月21日）.
[4] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html（オンライン）（閲覧日：2015年 9 月14日）.
[5] 奥村学（監修），佐藤一誠．トピックモデルによる統計的潜在意味解析，コロナ社，2015.
[6] 奥村学（監修），高村大也．言語処理のための機械学習入門，コロナ社，2010.
[7] http://mallet.cs.umass.edu/index.php（オンライン）（閲覧日：2015年 9 月14日）.
[8] http://svmlight.joachims.org/（オンライン）（閲覧日：2015年 9 月14日）.

[9] Zelizer, B. Taking Journalism Seriously: News and the Academy, SAGE Publications, Inc, 2004.
[10] クラウス・クリッペンドルフ（著），三上俊治（翻訳），橋元良明（翻訳），椎野信雄（翻訳）．メッセージ分析の技法―「内容分析」への招待，勁草書房，1989.
[11] 上野栄一．内容分析とは何か：内容分析の歴史と方法について，福井大学医学部研究雑誌第９巻第１号・第２号合併号，pp.1-18，2008.
[12] 工藤文．中国における批判報道の特性―新聞の内容分析を通して―，日本マス・コミュニケーション学会2011年度秋季研究発表会・研究発表論文，pp.1-6，2011.
[13] 中村理．報道への疑問をどのように研究へつなげるか？―内容分析の手法―，早稲田政治經濟學雜誌，No.387，pp.10-15，2015.
[14] 李光鎬．韓国の TV ニュースにおける日本関連報道の内容分析，慶應義塾大学メディア・コミュニケーション研究所紀要，No.57，pp.35-48，2007.
[15] 渡辺良智．新聞の東日本大震災報道，山學院女子短期大學紀要，No.65，pp.63-82，2011.
[16] 乾孝司，奥村学．テキストを対象とした評価情報の分析に関する研究動向，自然言語処理，Vol.13，No.3，pp.201-241，2006.
[17] Kanda, R., S. Tsuji, and H. Yonehara. Text Mining Analysis of Radiological Information from Newspapers as Compared with Social Media on the Fukushima Nuclear Power Plant Accident, Journal of Disaster Research, Vol.9, No.7 pp.690-698, 2014.
[18] Miller, M. M. and B. P. Riechert. The Spiral of Opportunity and Frame Resonance: Mapping the Issue Cycle in News and Public Discourse; Framing Public Life: Perspectives on Media and Our Understanding of the Social World, Routledge, editer Reese, S. D., O. H. Gandy Jr., and A. E. Grant, pp.106-122, 2001.
[19] 白井匡人，三浦孝夫．LDA を用いた著者推定，DEIM Forum 2011 F4-3，2011.
[20] 高須淳宏，相原健郎．テキスト分類における訓練データと性能の実験的考察（（特集）電子文書処理），NII journal 6，pp.1-8，2003.
[21] 吉岡康平，小枝正直．BM25を用いた関連語抽出と単語分類，情報処理学会全国大会講演論文集，Vol.74，No.1，pp.553-554，2012.
[22] Blei, D. M., A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp.993-1022, 2003.
[23] 山本浩平，江口浩二，高須淳宏．カテゴリ階層を考慮した確率的トピックモデルのモデル選択付き学習，DEIM Forum 2012，F5-2, 2012
[24] http://www.slideshare.net/MasayukiIsobe/ss-35851169（オンライン）（閲覧日：2015年９月14日）．
[25] Rieck, K., S. Sonnenburg, S. Mika, C. Schäfer, P. Laskov, D. Tax, and K.-R. Müller. Support Vector Machines; Handbook of Computational Statistics, editor Gentle, J. E., W. K. Härdle, and Y. Mori, p883-926, 2012.
[26] Müller, K.-R., S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An Introduction to Kernel-Based Learning Algorithms, IEEE Transcations on Neural Networks,

Vol.12, No.2, pp.181-201, 2001.
[27] http://khc.sourceforge.net/（オンライン）（閲覧日：2015年９月14日）.
[28] Büttcher S., C. L. A. Clarke, and G. V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines, The MIT Press, 2010.
[29] 樋口耕一．社会調査のための計量テキスト分析―内容分析の継承と発展を目指して，ナカニシヤ出版，2014.
[30] 内田治．すぐわかる SPSS によるアンケートの多変量解析（第３版），東京図書，2011.
[31] 高村大也，乾孝司，奥村学．スピンモデルによる単語の感情極性抽出．情報処理学会論文誌，Vol.47 No.2，pp.627-637，2006.
[32] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一．意見抽出のための評価表現の収集，自然言語処理，Vol.12 No.3，pp.203-222，2005.
[33] 東山昌彦，乾健太郎，松本裕治．述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第14回年次大会論文集，pp.584-587，2008.
[34] 菅原久嗣，アレナ・ネビアロスカヤ，石塚満．日本語テキストからの感情抽出，The 23rd Annual Conference of the Japanese Society for Artificial Intelligence, pp.1-2, 2009.