テキストマイニングを用いた新聞メディアの報道傾向検出への試み

市川, 祐太 / ICHIKAWA, Yuta

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
57
(開始ページ / Start Page)
1
(終了ページ / End Page)
5
(発行年 / Year)
2016-03-24
(URL)
https://doi.org/10.15002/00013089

テキストマイニングを用いた 新聞メディアの報道傾向検出への試み

PRELIMINARY STUDY ON DETECTION OF NEWSPAPER TREND AMONG THEIR PUBLISHERS USING TEXT-MINING APPROACH

市川祐太
Yuta Ichikawa
指導教員　彌冨 仁

法政大学大学院理工学研究科応用情報工学専攻修士課程

News provided by mass media is required its accuracy, fairness and objectiveness. However, it is often biased by “hidden” thought of each media. In this paper, we challenged to detect such biases among newspaper articles provided three Japanese publishers; Asahi, Mainichi and Yomiuri newspapers. We collected online articles of those papers and Wikipedia as a supplemental data, and (1) divided them into set of each single words using morphological analysis, (2) obtained vector representation of each word by the word2vec methodology, (3) adjusted vector orientation based on the value registered on the Semantic orientation dictionary, and (4) compressed the vector dimension with t-SNE (t-distributed Stochastic Neighbor Embedding) and PCA(Principal Component Analysis) for data visualization. We objectively confirmed from our results that usage of words in news articles has not small variety among media.

１．はじめに

マスメディアには報道倫理に則り、受け手が誤って解釈をしてしまうことや、ある側面のみを作為的に強調することが無いように、客観的で正確性、公平性のある報道を行う必要がある。しかしながら、報道倫理の概念は法律などで明確な基準が定められているわけではなく、実際にはメディアやジャーナリストが自主的に設けているものであるため、それぞれに報道姿勢が大きく異なっている。つまり、マスメディアにおいて報道の偏りと言うものは少なからず存在する。

一方で情報の受け手は、 前述のような報道の偏りが存在していることを意識して情報を受け取らなければ、作為的な報道においても正確、 および公平なものであると受け取ってしまい、結果としてマスメディアによる大衆の刷り込みなどが起きてしまう可能性がある。 その対応策として、様々なメディアから情報を集め、それらを基に客観的な判断を行う事が挙げられるが、受け手が積極的に情報収集を行わなければならないため、非常に手間がかかる。

そこで本研究では、マスメディアから発信されている情報から、メディア毎の報道の偏りを検知し、可視化することで、受け手が客観的な判断を行うための手助けとなるようなシステムの構築を行う事を最終目標とした。

本研究のシステムを構築するにあたり、大量の文章データに対して解析を行い、その中に潜んでいる相関関係やパターンなどを知識として取り出すテキストマイニングと呼ばれる手法を用いる。テキストマイニング技術の応用例として、近代日本小説家の文章の書き方による著者の自動判別[1]や、文書からのトピック抽出[2]など様々な研究が行われている。

本研究と類似する先行研究として、熊本らによる Web ニュース記事からの喜怒哀楽抽出[3]があり、Web ニュースサイトの記事に対して受け手がどのように感じるかを推測するシステムを構築している。しかしながら、このシステムは受け手それぞれに対しての主観的な感情の適応を目指しているため、 本研究とはアプローチが異なる。

また、田中らによる公共政策に関する大手新聞社説の論調についての定量的物語分析[4]では、大手新聞社の社説を対象にした定量的な分類が行われているが、本研究ではメディアの意見が直接表に出る社説ではなく、記事の客観性が求められる社会面や政治面における、新聞社において同じカテゴリの記事での報道の偏りを検知する

ことを目標とする。

２．対象と方法

本研究の最終目標は、マスメディアから発信されている報道の偏りを検知し、どのように偏りがあるか可視化するシステムの作成である。

特に新聞社に着目すると報道媒体は新聞記事であり、文章である。そこで、本研究では更に文章を単語に分解し、単語の使われ方によって各新聞社における単語の傾向を検知出来るかを検討した。

Web 新聞社３社と日本語 Wikipedia の記事を対象に、Word2Vec[5]を用いることで単語ベクトル表現を獲得し、感情極性値[6]を付与することで、各新聞社の単語の使われ方の違いを検出した。

本研究における処理の流れを図１に示す。

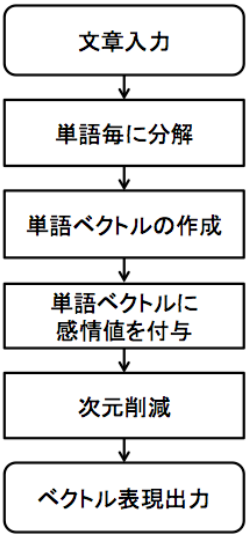

図 1.処理の流れ

（１）取り扱う文章データ

解析を行うには大量の文章データの取得が必要である。そこで、解析対象を朝日新聞デジタル[7](以降朝日と表記)、毎日新聞[8](以降毎日と表記)、読売オンライン[9](以降読売と表記)の３社から、社会、政治、経済、国際カテゴリの記事を取得した。

また、本システムの手法として後述する Word2Vec を用いて単語表現モデルを獲得して解析を行った。その際にいずれかの新聞社で用いられていない単語が存在すると、それぞれの新聞社で単語ベクトルの比較が行えないため、Wikipedia 日本語版コーパス[10](以降 Wiki と表記)の記事データを新聞社３社の記事と合わせて Word2Vec の処理を行った。また、Wiki のみにおいても比較のため同様の行うこととし、次の 4 つのデータセットを用いた。

- Wiki 記事群
- Wiki 記事群 + 朝日記事群
- Wiki 記事群 + 毎日記事群
- Wiki 記事群 + 読売記事群

（２）単語毎に分解

本システムでは、文章を単語に分けた上で、その単語が各新聞社でどのような使われ方の違いがあるかを判別する。入力された文章を単語ごとに分解するために、単語単位に分かち書きを行うことが出来る Mecab[11]という形態素解析ツールを用いた。これは、文章を入力すると IPA 辞書と呼ばれる独自の辞書によって、文章中の単語の品詞やその基本形、活用型などといった情報を含めて解析を行うことが出来るオープンソースのツールである。

よって、本システムでは文章を単語に分解した上で、動詞や形容詞は基本形に直して解析を行った。また、分かち書きの精度を上げるため、Mecab による形態素解析の後に 2 点の処理を行った。

A) 流動的な単語への対応

形態素解析の大きな課題の１つとして、人物名や新しい単語など、流動的な単語の解析が困難な事が挙げられる。Mecab においても、IPA 辞書に予め登録されている単語と、入力された文章を照らし合わせていくことで形態素解析を行っているため、IPA 辞書に登録されていない新しい単語や人物名などの単語は正確に単語に分けることが出来ない。

新聞や Wikipedia における文章には、歴史上の人物や著名人などの名前の他にも新しい単語などが多く存在するため、それらを正確に形態素解析する事が望ましい。

そこで、本システムでは Mecab の標準辞書である IPA 辞書に加え、佐藤らが作成した新語辞書[12]を用いることで、形態素解析の精度の改善を試みた。

B) 否定の意味を持つ助動詞への対応

流動的な単語に加えて形態素解析の課題となるのが、否定の意味を持つ助動詞である。否定が持つ意味は重要な文脈において要素である。

本システムでは、単語に感情極性値と呼ばれる値を割り振る。この感情極性値は、その単語がポジティブな（良い）印象を持っているか、ネガティブな（良くない）印象を持っているかを表す値である。

しかしながら、Mecab による形態素解析では「良くない」を「良い」と「ない」に変換してしまうため、「良くない」の単語モデルが正確に表現することが出来ない。よって、動詞や名詞の直後に存在する否定の意味の助動詞「ない」および「ぬ」に対して、前の単語と連結させて１単語とする処理を行った。

（３）単語ベクトルの作成

品詞分解された単語群を用いて、それぞれの単語について、任意次元の単語ベクトル表現を獲得した。単語ベクトルを作成する際には Word2Vec[5]と呼ばれる手法を用いた。

Word2Vec は Mikoliv らによって提案された、”同じ文脈で利用される単語は、同じ意味を持つ”という仮説に基づき、単語をニューラルネットワークにより学習し、ベクトルで表現する技術である。単語の特徴や意味構造を含んでおり、かつベクトル表現であるために、単語間での距離の計算によって類義語の抽出や単語同士の加算減算が可能になる。これにより、本研究では各新聞社の単語モデルを獲得した後、それぞれの新聞社でどのように単語の特徴や意味構造が異なるのかの検証を行った。

Word2Vec の入力として、形態素解析が行われて単語に分かち書きされている文章を用いる。また、注目する単語において、文脈として前後 n 単語を対象とするかを定める。その範囲のことをウィンドウサイズと呼ぶ。

本研究においては、Word2Vec の学習モデルに Skip-gram モデルと呼ばれる手法を用い、ウィンドウサイズを 5 として解析を行った。

ウィンドウサイズ $n=5$ とした時の処理の流れを図 2 に示す。初めに、入力された文章に対して 1-of-K 表現を作成する。ここで、入力文章の総語彙数を m とする。

1-of-K 表現とはある要素が 1 であり、それ以外の要素が 0 であるような表現方法である。ここでは注目単語 t の要素が 1 であり、それ以外の要素が 0 である注目単語の 1-of-K 表現 $w(t)$が入力層に与えられる。

入力層から中間層への重みを W^{XH}、中間層から出力層への重みをとすると W^{XH}は $m \times d$行列、W^{HY}は $d \times m$行列となる。

次に、中間層へ $w(t)$に重み W^{XH}を掛けた結果が出力される。つまり、中間層では注目単語 t における d次元のベクトル表現が得られる。最後に、出力層へ d次元の単語ベクトル表現に重み W^{HY}を掛けた結果が出力される。

つまり、m次元の 1-of-K 表現が得られる。この出力を、$w(t-5)$から $w(t+5)$の値に近づくように重みを更新するように学習が行われる。この処理を繰り返し、すべての単語において d 次元の表現ベクトルを得た。

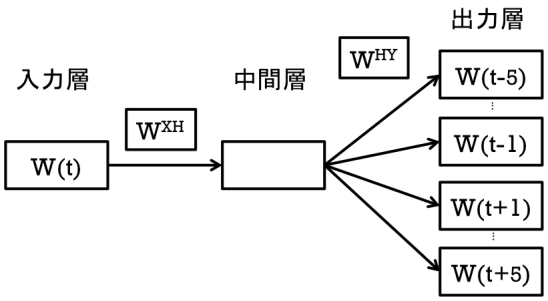

図 2.Skip-gram モデル

（４）単語ベクトルに感情符号を付与

Word2Vec により獲得した各新聞社の単語モデルは、周辺単語の出現確率を基に学習を行っているため、 文脈による特徴は捉えているが、その単語自体の内容の特徴は捉えていない。そこで、高村らが作成した単語感情極性辞書[6]を用いて、単語自体の特徴を付与することを試みた。

感情極性辞書とは、ある単語と、その単語が一般的に良い印象を持つか(positive)、悪い印象を持つか(negative)を 1 から-1 までの値で表した対応表である。この対応表に入っている単語は、岩波国語辞書をリソースとしており、約 5 万単語に感情極性値が与えられている。本システムにおいて、この対応表に当てはまる単語について感情極性値に付与されている符号(+か-)を重みとして、その単語ベクトルの値を更新した。これにより、周辺単語の情報に加え、良い印象を持つ単語か悪い印象を持つ単語か、といった情報を付与出来ると期待した。

（５）次元削減

報道の偏りの可視化を目的としている本研究において、どのように単語および文章の傾向を視覚的に表現するかは極めて重要である。そこで、本研究では t-SNE(t-distributed Stochastic Neighbor Embedding)[13]と主成分分析(PCA)の 2 種類の次元削減法を用いて、それぞれ視覚的にどのように異なっているかを主観的に判断した。t-SNE は Laurens らによって提案された次元削減法であり、高次元の点間の距離と、次元圧縮後の点間の距離を、勾配法を用いて最小化する手法である。つまり、元の特徴空間距離の比率を保ったまま、次元を削減することが出来る。

（６）ベクトル表現出力

t-SNE と主成分分析によって次元を削減し、それぞれ出力として可視化し、考察を行った。本システムにおいて、新聞社による単語の使われ方がどのように違うか、また視覚的にどう異なるかを検証した。

３． 結果および考察

本システムに入力する前の取得したデータ量について表 1 に示す。新聞社の記事データがそのままでは少ないため、本システムに入力する際には、傾向をより顕著にする目的で新聞社の記事データを 10 倍に増加させる処理を行った。前述の 4 つのデータセットを入力として、表 2 に示すようなパラメータを用いて様々なパターンの可視化を実現した。

表 1.取得した記事データ

データ名	データサイズ(MB)
Wikipedia 日本語	5270
朝日新聞デジタル	107
毎日新聞	40
読売オンライン	86

表 2.解析パラメータ一覧

単語モデルの次元数 d	10, 25, 50, 100, 200
感情極性辞書	付与なし, 感情符号
次元削減法	t-SNE or PCA

（１）Word2Vec で生成したベクトルの次元数による変化

はじめに、全てのデータセットにおいて共通している文章である Wiki について検証を行った。単語モデルの次元数それぞれにおいて、主成分分析にて 2 次元へ削減し、寄与率の違いを検証した。

その結果を表 3 に示す。

表 3. 単語モデルの次元数ごとの削減後の寄与率

	感情付与なし(%)	感情符号(%)
次元数 10	49.56	80.33
次元数 25	31.32	67.63
次元数 50	23.90	58.57
次元数 100	19.43	48.96
次元数 200	15.81%	39.97%

どの次元数においても感情付与なしのパターンでは寄与率が低い。また次元数を下げると、全ての場合において寄与率が上昇していることが分かる。これは、主成分分析を行う際の削減する次元数自体が少なくなるからであると推測できる。この結果より、主成分分析における可視化において次元数 10 のモデルを用いれば情報欠損が少ないと考えられる。

しかし、Word2Vec の段階にて次元数を少なくしている事から、単語モデル自体の精度の低下が考えられる。よって、同じ条件下における単語モデルの精度についても検証を行った。

（２）単語モデルの次元数における精度

直感的に理解しやすい単語において、Word2Vec の次元数による精度について検証を行った。ここでは感情付与なしの Wiki データセットにおいて対義語が既知な形容詞単語を主観的に 5 個選択し、その対義語とのコサイン類似度を求めた。その結果を図 3 に示す。

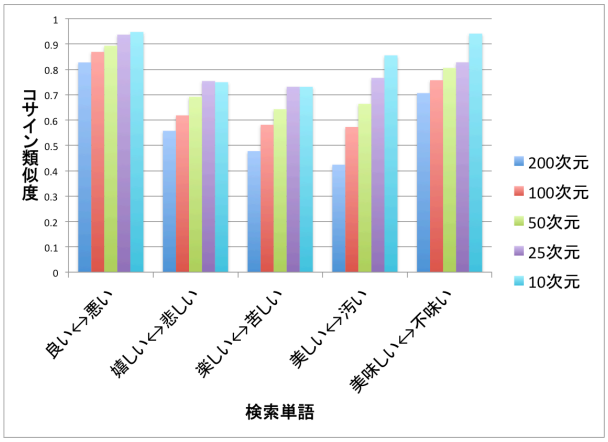

図 3.各次元における既知な対義語のコサイン類似度

この結果においては感情付与なしであることから、対義語の単語に対するコサイン類似度は高いほど精度が良い事に留意する。次元数を少なくすると、コサイン類似度も上昇しているものが多いことが分かる。しかし、「嬉しい⇔悲しい」、「楽しい⇔苦しい」においては次元数 25 と d=10 のコサイン類似度が並行または少し下がっている。この事から、次元数 25 から次元数 10 へモデルの次元数を下げる事による精度の低下が考えられる。

また、コサイン類似度が次元数 10 においても高くなっている「良い⇔悪い」について、次元数 25 と次元数 10 において「良い」に最もコサイン類似度が近い単語を求めた。その結果を表 4 に示す。

表 4. 次元数 25 と 10 における「良い」の上位 5 単語

	25 次元	10 次元
1 位	よいない	丁寧
2 位	とても	雰囲気
3 位	悪い	あまりに
4 位	作れるない	表情
5 位	合うない	それでいて

次元数 25 において、「悪い」や、「よい」の否定語である「よいない」が上がっている。対して次元数 10 では、「悪い」が挙がっていない。この事から、次元数 10 における単語ベクトルの表現の精度が 25 次元に比べると低いと考えられる。よって本研究では、次元削減での可視化について、Word2Vec における単語表現が主観的に成り立っており、かつ寄与率 67%を実現している次元数 25 が良いと過程して可視化を行った。

（３）主成分分析と t-SNE による可視化比較

各新聞社における単語単位での傾向の違いを検証するため、次元数 25 による感情符号付与後の Word2Vec モデルから、主成分分析と t-SNE による 2 パターンの次元削減を用いて、それぞれの単語モデルを 2 次元で可視化を行った。このとき、全てを出力すると単語数が多く比較が困難なため、形容詞や名詞など品詞に分け、その中でも主観的に単語の印象が分かりやすい単語を 100 個ずつ抽出することで可視化を行った。形容詞における各新聞社を含んだデータセットについて、t-SNE の 2 つを用いて可視化した結果を、「良い」の単語の周辺で拡大したものを図 4, 5, 6 に示す。

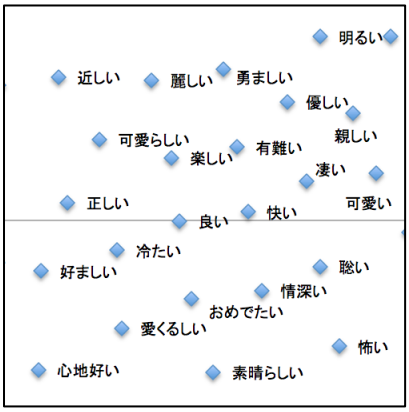

図 4. Wiki+朝日データセットにおける t-SNE 結果

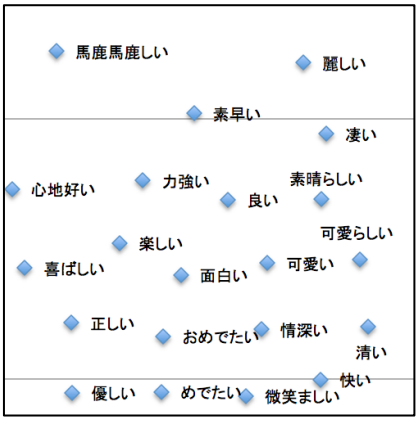

図 5. Wiki+毎日データセットにおける t-SNE 結果

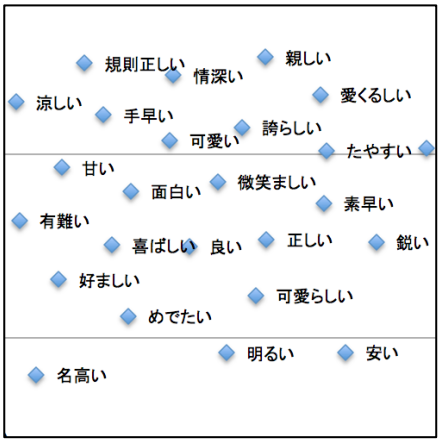

図 6. Wiki+読売データセットにおける t-SNE 結果

ここで、可視化における軸の留意点について述べる。それぞれのデータセットにおける単語モデル次元数は同じであるが、入力文章の 1-of-K 表現自体が異なるため、データセットを跨いだ直接的な単語ベクトルの比較は行うことが出来ない。また同様に、その単語ベクトルを次元削減したものについても同様であり x 軸、y 軸の意味合いはそれぞれのデータセットによって異なる。

t-SNE による次元削減法では、高次元空間による点同士の距離比率を保ったまま可視化することが出来る。

図 6 から、読売新聞における「良い」と距離が最も近い単語に「喜ばしい」が存在する。つまり、「良い」と「喜ばしい」は読売新聞において似たような文脈で、似たような感情を持つ単語として用いられていると推測できる。この事から「喜ばしい」と言った単語においては、読売新聞は他の新聞社と比較すると「良い」に近い意味合いとして用いられているのではないかと推測でき、新聞社独自の傾向と考えられる。また、主成分分析における可視化では、上記のような周辺単語の違いが顕著に現れず、推測することが出来なかった。

4. まとめ

新聞社 3 社を対象に、Word2Vec を用いて単語モデルを獲得し、感情値を付与して可視化を行った。その結果、各新聞社で記事中の形容詞単語の傾向が主観的ではあるが推測できた。今後は、評価指標の構築行い、文章、記事へと解析の幅を広げていくことで、新聞社の傾向検出が可能になると考えられる。

謝辞：本研究にあたり、全般にわたるご指導をしてくださった彌冨仁准教授、および彌冨研究室の皆様に深く御礼申し上げます。

参考文献

1) 松浦司, 金田康正：近代日本小説家8人による文章の n-gram分布を用いた著者判別, 情報処理学会研究報告. 自然言語処理研究会報告, Vol.53, pp.1-8, 2000.
2) 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道：文書クラスタリングによるトピック抽出および課題発見, 社会技術研究論文集. Vol.5, pp.216-226, 2008.
3) 熊本忠彦, 田中克己：Web ニュース記事からの喜怒哀楽抽出, 情報処理学会研究報告. 自然言語処理研究会報告 Vol.1, pp.15-20, 2005.
4) 田中皓介, 中野剛志, 藤井聡：公共政策に関する大手新聞社説の論調についての定量的物語分析, 土木学会論文集. Vol.69, pp.353-361, 2013.
5) Tomas Mikolov, Kai Chen, Greg Corrade, JeyDean : Ecient Estimation of Word Rep-resentations in Vector Space, 2013
6) 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ジャーナル, Vol.47 No.02 pp.627-637, 2006.
7) 朝日新聞デジタル:” http://www.asahi.com/”
8) 毎日新聞:” http://mainichi.jp/ “
9) 読売オンライン:” http://www.yomiuri.co.jp/”
10) Wikipedia 日本語コーパス: “https://dumps.wikimedia.org/jawiki/”
11) Taku Kudo:MeCab:Yet Another Part-of-Speech and MorphologicalAnalyzer, ” http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html”
12) Toshinori Sato : Neologism dictionary based on the language resources on the Web for Mecab, ” https://github.com/neologd/mecab-ipadic-neologd/”, 2015
13) Laurens van der Maaten , Geoffrey Hinton : Visualizing Data using t-SNE, 2008