機械学習と統計的検定を利用した知見獲得とその評価

董 卜睿 [*1]　村田 真樹 [*2]　馬 青 [*3]
[*1] 鳥取大学大学院 持続性社会創生科学研究科 工学専攻
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*3] 龍谷大学 先端理工学部 数理・情報科学課程
[*1,*2]{m20j4163a@edu.,murata@}tottori-u.ac.jp
[*3]qma@math.ryukoku.ac.jp

概要

本研究では，最大エントロピー法，BERT，SVM の 3 つの教師あり機械学習法と，符号検定，区間推定の 2 つの統計手法を用い，新聞データから単語対の情報を収集し，知見を獲得する．獲得した情報 (知見) を関連性と有益性に基づき評価する．

関連性評価では，12 年分の新聞データで訓練した Word2vec から得た，事前に人手で設定した単語対 (テーマキーワード対) に最も類似する 500 個の単語と，人手で単語対について連想した 30 個の単語を正解として，提案手法によって得られた最も評価の高い 500 個の単語との一致率を計算した．計算の結果，統計的手法が機械学習法を上回った．最も一致率が高かったのは符号検定であり，Word2vec との一致率は 0.126，人手で連想した単語との一致率は 0.344 であった．有益性評価は，提案手法で最も評価の高い 100 個の単語を取り出し，被験者がどの手法から取得したのかを分からないようにして，単語の有益性を評価する．全新聞 (2007 年から 2018 年までの全て) を訓練した Word2vec が最も評価が高く，平均の比率は 0.594 である．提案手法の中で最も評価の高い手法は BERT であり，平均の比率は 0.444 となった．

1　はじめに

メディアの電子化により，多くの電子テキストが出現し，そこから重要な情報を素早く入手することが求められている．

そこで，教師あり機械学習と統計的検定を利用して，新聞データから様々な分野に関するテーマキーワード対の知見を獲得することを本研究の目的とする．本研究では，教師あり機械学習と統計的検定で取り出した単語を知見とする．得られた知見の評価に関連性評価と有益性評価を用いる．なお本研究で用いる単語対は「健康」と「病気」,「政治」と「経済」,「輸入」と「輸出」,「社会主義」と「資本主義」,「オリンピック」と「パラリンピック」である．この単語対をテーマキーワード対と呼ぶ．単語対の一方の単語はテーマキーワードと呼ぶ．

2　先行研究

村田らの研究 [1] はウェブと新聞から株式相場における情報を収集している．パターンと教師あり機械学習を利用して株式相場や経済に関わる様々な知見獲得を行い，獲得した知見は日本の株式相場の日々の変動を予測するために使用される．

Bollen ら [2] は，Twitter のメッセージを用いて，大統領選挙と 2008 年の感謝祭に対する人々の感情的な反応を分析し，ダウ平均の日々の上昇と下降を予測した．

鎌倉 [3] は，多くのテーマキーワード対において，機械学習の最大エントロピー法と統計的検定の符号検定により知見を獲得した．獲得した知見の有効性を人手で評価した．本研究に使用した手法は，鎌倉が使用した 2 つの手法以外に，提案手法として SVM，BERT と区間推定を追加した．また，比較手法として Word2vec を追加している．

3　提案手法

テーマキーワード対を含む文章から得られる単語を素性とする．本研究では, 教師あり機械学習と統計的手法を利用する．2007 年から 2018 年の毎日新聞テキストデータからテーマキーワード対に関わる素性を分析し，複数のテーマキーワード対の知見を収集する. 手法ごとに得られた素性が人にとって役

に立つのかを明らかにするために人手による評価を行う．

3.1 素性分析

テーマキーワード対をA, Bとして以下に提案手法を示す. 本研究では文章として，新聞の一段落を利用する．

- 手順１：新聞データよりＡまたはＢを含む文章 (段落) をそれぞれ収集する．
- 手順２：A, Bを消して，Xに置き換える．
- 手順３：手順１で収集したＡとＢに関する文章 (段落) をランダムに整列して，単語ＡとＢを公平に分析するために，ＡとＢを含む文章 (段落) を同数にする．
- 手順４：手法を利用して素性を抽出する．
- 手順５：手法ごとに得られた素性の上位 500 個を取り出す．
- 手順６：手順５で得られた素性を評価する．

3.2 最大エントロピー法

最大エントロピー法 [1, 3, 4] は，どのテーマキーワードが記事中に出現する可能性が高いかを学習できる．テーマキーワード A , B があるとする．A または B を含む段落を収集する．収集した段落から A と B を取り除き，X とする．学習結果に基づいて，X にテーマキーワード A または B のどちらがあったかを推定する．学習段階で，素性に対するテーマキーワードの重要度を示す正規化 α 値を計算し，上位 500 個の素性を分析する．

3.3 SVM

SVM[5] とは，support-vector machine の略で，空間を超平面で分割することにより 2 つの分類からなるデータを分類する手法である．2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負例の間隔 (マージン）が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ，このマージンを最大にする超平面を求めそれを用いて分類を行う．テーマキーワード対 A,B があるとする．A または B を含む段落を収集し，その段落から A と B を取り除き，X とする．その段落と X に入るものの対を学習データとして超平面を学習する．学習で得られたモデルに基づいて，単語 X が平面に対してどのあたりにあるかを計算し，X が A または B であるかを推定する．素性分析する方法は以下である．段落を個々の単語に切り分け，学習済み svm モデルを用いて各単語について分離平面までの距離を計算する．算出された距離は，それが該当するエリアとしてプラスマイナスで表現される．分離平面との距離が大きいほど有用な素性である．上位から 500 個の素性を抽出し，分析する．

3.4 BERT

BERT とは，Bidirectional Encoder Representations from Transformers の略で，Jacob Devlin ら [6] の論文で発表された自然言語処理モデルである. ラベル付けされていないテキストから，全層で左右両方の文脈を共同で条件付けすることにより学習する．BERT は双方向の Tranceformer によって学習を行う．

テーマキーワード対 A,B があるとする．A または B を含む段落を収集し，その段落から A と B を取り除き，X とする．学習したモデルに基づいて，X を推定する．素性分析の方法は以下である．段落を個々の単語に切り分け，学習済み BERT モデルを用いて各単語についてテーマキーワード対の確率値を計算し，上位 500 個の単語を有用な素性として抽出し分析する．

3.5 符号検定

符号検定 [3] は，勝ちと負けや表と裏といった相反する 2 つのペアの差に符号 + と - を割り当てる統計的検定として用いられる．二項定理に基づく片側検定により， A または B を含む段落の中で A と B と伴って出現する単語が，いずれのカテゴリーにおいても，全データでの出現率よりも高い頻度で出現するかどうかを判定し，有意確率 p 値を求める．p 値が低いほど有用な素性であるので，本研究は p 値の下位 500 個の素性を取り出し，分析する．

3.6 区間推定

テーマキーワード対を A,B とする，新聞データから A または B を含む段落を収集して，A と B の出現回数を数える．A の出現回数は N_a 回で，B の出現回数は N_b 回，A の真の出現率は θ とする．$N = N_a + N_b$ とする．実際の A の出現率は $\frac{N_a}{N}$ により求められる．「真の A の出現率 θ は概率 95 % で $a \leq \theta \leq b$ の区間にある」と考える．このような考えを区間推定という．下限値 a 値が高いほど有用な素性である．上位の 500 個の素性を抽出する．

表 1　全評価データ

素性対	段落数
健康, 病気	28,086
オリンピック, パラリンピック	16,574
輸入, 輸出	27,852
社会主義, 資本主義	2,338
政治, 経済	40,000(155,776)

表 2　単語「健康」の人手連想評価の例

連想単語	ME	BERT	区間	符号	SVM
運動	運動	おいし	増進	増進	マスク
野菜	野菜	治る	野菜	管理	代謝

表 3　単語「病気」の人手連想評価の例

連想単語	ME	BERT	区間	符号	SVM
病院	失業	失明	入退院	病院	武力
医者	療養	損傷	死	死	主治医

4　評価

4.1　評価データ

評価には，2007 年から 2018 年の毎日新聞を用いて，評価データの内訳を表 1 に示す．テーマキーワード対「政治」「経済」の段落数は 155,776 であるが，機械学習手法は 15 万データを訓練すると，時間がかかるので，最大エントロピー法，SVM，BERT，符号検定と区間推定の評価データはランダムに抽出した 40,000 個の段落を用いる．最大エントロピー法と SVM は学習データとテストデータの比率は 1 対 1，BERT の訓練データ，検証データとテストデータの比率は 3 対 1 対 4 となるように全評価データを分割して用いる．最大エントロピー法と SVM と BERT は，訓練データのみから素性を取り出す．(最大エントロピー法では訓練データで学習する正規化 α 値を用いる．SVM と BERT は，訓練データの段落にあった個々の単語を入力として用いることで素性分析する．) 符号検定と区間推定は表 1 の全評価データを使用した．Word2vec は 2007 年から 2018 年まで全ての新聞を訓練する．

4.2　関連性評価：Word2vec 自動評価

Word2vec とは，大量の電子テキストを解析し，各単語の意味をベクトル表現することができる手法である．単語をベクトル化することで，単語の意味の近さを計算し，それを使って似たような意味の単語を探すことができる．Word2vec で全新聞 (2007 年から 2018 年までの全て) を訓練し，テーマキーワードと類似する単語を探し，類似度の高い 500 語を正解として提案手法を評価する評価を行う．

- 手順 1 ：Word2vec に評価データを学習させる．
- 手順 2 ：手順 1 で作成したモデルにテーマキーワードを入力し，テーマキーワードと類似する単語を求める．
- 手順 3 ：テーマキーワードとの類似度が最も高い 500 個の単語を抽出し，3.1 節で収集した素性との一致数を数える．

4.3　関連性評価：人手連想評価

3.1 節で収集したテーマキーワード対の手法ごとの素性を評価対象とする．人手でテーマキーワード対から連想する単語を 30 個書き出し，3.1 節で収集した素性 500 個との一致数を数える．Word2vec は，全部の新聞データを学習する．提案手法と比較するために使用される．表 2 と表 3 は連想評価の例である．人手で連想した単語と手法で取り出した上位の素性が一致したときの数を数える．人手連想評価の手順は以下のとおりである．

- 手順 1 ：人手でテーマキーワード対から連想する単語を 30 個書き出す．
- 手順 2 ：手順 1 で収集した連想語 30 個と 3.1 節で収集した素性 500 個との一致数を数える．

4.4　有益性評価

収集した素性がどの方法で得られたものかを被験者がわからないようにするために，各手法から収集した素性を合わせて，意味ソート [7] を行う．そして人手で評価する．意味ソートでは，単語を意味順に並べることができる．似たような意味を持つ単語を近くに配置することで，手作業によるチェックを効率的に行うことができる．表 4 と表 5 に有益性評価の例を示す．

Word2vec は全部の新聞データを学習する．提案手法と比較するために使用される．

- 手順 1 ：各手法 (最大エントロピー法，BERT，SVM，符号検定，区間推定，Word2vec) の素性分析の上位 100 個を取り出す．
- 手順 2 ：手順 1 で得られた素性をすべて合わせて，意味ソートを行う．
- 手順 3 ：被験者がどの手法によりその素性が得られたかを知らない状況で評価する．

評価基準は以下で示す．

表 4　「健康」の有益性評価の例

手法に取り出す単語	評価
養生	○
予防	○
医薬	△

表 5　「病気」の有益性評価の例

手法に取り出す単語	評価
感染	○
移植	△
オーチャード	×

表 6　機械学習手法の正解率

テーマキーワード対	ME	SVM	BERT
健康, 病気	0.8363	0.8187	0.9711
オリンピック, パラリンピック	0.9118	0.9463	0.9599
輸出, 輸入	0.8325	0.8185	0.9210
政治経済	0.8707	0.8338	0.9525
社会主義, 資本主義	0.7979	0.7851	0.8441
平均	0.8498	0.8405	0.9297

- ○：テーマキーワードの関連語であり，被験者はこの素性に興味を持ち，被験者が役に立つと思う単語．
- △：テーマキーワードの関連語であり，被験者はこの素性に興味を持たず，被験者が役に立たないと思う単語．
- ×：テーマキーワードの関連語でない，または単語の意味がわからない．

5　評価結果

表 6 は，新聞記事のテストデータでの機械学習手法の正解率を示している．表のように，BERT は，他の機械学習手法よりも正解率が高く，平均正解率は 0.9297 であった．

表 7 は，4.2 節の評価の全てのテーマキーワード対の平均である．Word2vec 自動評価は，各手法で取り出した単語数は 500 個なので，一致の数を 500 で割った比率で表示する．統計的手法が機械学習手法よりも優れていることがわかる．平均一致率は 0.126 であった．機械学習の結果は，BERT，ME，SVM の順で一致率が高くなった．

表 8 は，4.3 節の評価結果である．4 人の被験者がテーマキーワード対に関して連想した語句 30 個と五種類の手法で得た素性各 500 個との一致率を求めた．人手で連想した単語数は 30 なので，4 人の一致数の平均を 30 で割った比率で表示した．Word2vec 自動評価の結果と同様に，統計的手法は機械学習手法を上回った．最も性能が良いのは符号検定であり，平均一致率は 0.344 である．

表 7　関連性評価:Word2vec 自動評価結果

単語	ME	SVM	BERT	符号	区間
平均	0.071	0.07	0.081	0.126	0.113

表 8　関連性評価:連想評価結果

単語	ME	SVM	BERT	符号	区間	Word2vec
平均値	0.233	0.202	0.135	0.344	0.257	0.296

表 9　有益性評価結果

	ME	SVM	BERT	符号	区間	Word 2vec
平均○	0.329	0.274	0.444	0.384	0.406	0.594
平均○+△	0.506	0.440	0.708	0.594	0.633	0.762

表 9 は 4.4 節の有益性評価の評価結果である．提案手法と Word2vec から各手法の評価の高い 100 個の単語を取り出し，1 つのテーマキーワード対で 1,200 個の単語 (100 × 2 単語× 6 手法) が評価される．被験者がどの手法によりその素性が得られたかを知らない状況で評価する．この評価の被験者は一人である．表 9 では，○は，被験者が有用と考える単語の比率である．また，○＋△は，被験者がテーマキーワード対に関連すると考えた単語の比率である．5 つのテーマキーワード対で 1 つの手法で取り出す単語の総数は 1,000 個なので，表 9 では 1,000 で割った比率で表示される．有益性評価の結果は，全ての新聞を訓練した Word2vec が最も評価が高く，平均の比率は 0.594 である．提案手法の中で一番良い手法は BERT で平均の比率は 0.444 となった．今後の課題として被験者を増やしたい．

6　おわりに

本研究では，3 つの教師あり機械学習法と，2 つの統計手法を用い，新聞データからテーマキーワード対の情報を収集し，関連性評価と有益性評価をする．関連性の評価では，符号検定が最も良い性能であり，性能は 0.344 であった．3 つの単語を連想したときに一つを推測できる割合である．また，より精度を高めるために，手法を組み合わせることは今後の課題としたい．有益性評価では，Word2vec が最も評価が高く，平均の比率は 0.594 である．有益性については，提案手法の中で最も評価が高かったのは BERT で，有用単語の比率は 0.444 である．有用単語の比率を増やす改良は，今後の課題としたい．

参考文献

[1]　村田真樹, 中原裕人, 馬青. パターンと教師あり機械学習と素性分析を利用したウェブと新聞か

らの株式相場に関わる知見獲得. 第 82 回全国大会講演論文集, Vol. 2020, No. 1, pp. 55–56, feb 2020.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8, 2011.

[3] 鎌倉周平. 機械学習と統計的検定を利用した知見獲得の評価実験. 鳥取大学卒業研究発表会論文, 2020.

[4] Masaki Murata, Kiyotaka Uchimoto, Masao Utiyama, Qing Ma, Ryo Nishimura, Yasuhiko Watanabe, Kouichi Doi, and Kentaro Torisawa. Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction. *Cognitive Computation*, Vol. 2, pp. 272–279, 2010.

[5] 村田真樹. 機械学習手法を用いた日本語格解析 : 教師信号借用型と非借用型, さらには併用型. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 101, No. 190, pp. 15–22, jul 2001.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, Vol. abs/1810.04805, , 2018.

[7] 村田真樹, 神崎享子, 内元清貴, 馬青, 井佐原均. 意味ソートｍｓｏｒｔ　意味的並べかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例:意味的並べかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例. 自然言語処理, Vol. 7, No. 1, pp. 51–66, 2000.

A　付録

A.1　意味ソート

単語を意味に応じて並べかえるという考え方は，Msort(meaning sort)[7] と呼ばれている．この意味の並べ替えの順番は，単語の羅列を表示する際には５０音順 (もしくはＥＵＣ漢字コード順) ではなく，単語の意味の順番に表示する．

単語を意味ソートすると，単語に対して意味的な順番で分類する．意味ソートで採用した分類は分類語彙表 (国立国語研究所 1964) に基づいている．動物，人間，組織，植物，生物の部分，自然物，生産物，空間，現象名詞，動作，精神，性質，関係，言語作品，その他，時間と数量の分類があり，全部で 17 種類の分類である．表 10 に意味ソートの例を示す．

表 10　意味ソートの例

分類	単語
(人間)	皇室 王室 官民
(組織)	全国 農村 県 日本 ソ連 寺 学校 学園 母校
(動作)	就任 まつり 祭り 祝い 巡礼 公式
(精神)	祝い 恒例 公式
(その他)	公式
(数量)	全国

表 11　Word2vec を用いた自動評価

単語	ME	SVM	BERT	符号	区間
健康	0.052	0.064	0.086	0.074	0.086
病気	0.084	0.082	0.224	0.166	0.176
オリンピック	0.058	0.066	0.046	0.140	0.160
パラリンピック	0.066	0.070	0.094	0.102	0.088
輸入	0.108	0.088	0.060	0.202	0.184
輸出	0.126	0.104	0.062	0.224	0.152
政治	0.038	0.042	0.030	0.068	0.050
経済	0.090	0.076	0.076	0.164	0.110
社会主義	0.062	0.068	0.116	0.066	0.066
資本主義	0.028	0.042	0.022	0.052	0.054
平均	0.071	0.070	0.081	0.126	0.113

A.2　評価結果

表 11 は，4.2 節の Word2vec 自動評価の結果である．表 7 に同様の表があるが，表 7 と違ってテーマキーワードごとの数値結果も表示している．Word2vec 自動評価は，各手法で取り出し単語数は 500 個なので，一致の数を 500 で割った比率で表示される．

表 12 は，4.3 節の連想評価の結果である．表 8 に同様の表があるが，表 8 と違ってテーマキーワードごとの数値結果も表示している．人手で連想した単語数は 30 なので，4 人の一致数平均を 30 で割った比率で表示．

表 13 は，4.4 節の有益性評価の結果である．表 9 に同様の表があるが，表 9 と違ってテーマキーワードごとの数値結果も表示している．提案手法と Word2vec から各手法の類似度の高い 100 個の単語を取り出すので，テーマキーワード対の数を 100 で割った比率で表示する．手法の平均数は 1,000(100 × 10 単語) で割った比率で表示する．

表 12　連想評価結果

単語	ME	SVM	BERT	符号	区間	Word 2vec
健康	0.192	0.183	0.175	0.267	0.208	0.250
病気	0.125	0.108	0.125	0.325	0.175	0.417
オリンピック	0.167	0.133	0.000	0.192	0.142	0.367
パラリンピック	0.308	0.317	0.267	0.408	0.217	0.358
輸入	0.125	0.108	0.000	0.192	0.100	0.292
輸出	0.067	0.042	0.017	0.175	0.092	0.300
政治	0.150	0.092	0.117	0.408	0.300	0.200
経済	0.408	0.367	0.292	0.633	0.492	0.433
社会主義	0.483	0.425	0.358	0.500	0.508	0.250
資本主義	0.300	0.242	0.000	0.342	0.333	0.092
平均値	0.233	0.202	0.135	0.344	0.257	0.296

表 13　有益性評価結果

	ME	SVM	BERT	符号	区間	Word 2vec
健康○	0.29	0.29	0.43	0.24	0.20	0.44
健康○+△	0.46	0.42	0.67	0.41	0.40	0.86
病気○	0.21	0.09	0.19	0.13	0.14	0.55
病気○+△	0.43	0.35	0.66	0.36	0.45	0.86
オリンピック○	0.24	0.27	0.31	0.38	0.41	0.62
オリンピック○+△	0.44	0.53	0.83	0.75	0.79	0.74
パラリンピック○	0.33	0.44	0.73	0.39	0.49	0.62
パラリンピック○+△	0.58	0.68	0.84	0.62	0.79	0.73
輸入○	0.24	0.33	0.35	0.40	0.29	0.27
輸入○+△	0.47	0.48	0.71	0.69	0.59	0.59
輸出○	0.43	0.31	0.34	0.38	0.34	0.53
輸出○+△	0.62	0.45	0.53	0.62	0.52	0.66
政治○	0.32	0.19	0.43	0.40	0.40	0.75
政治○+△	0.46	0.32	0.65	0.59	0.67	0.90
経済○	0.52	0.13	0.66	0.55	0.76	0.82
経済○+△	0.62	0.22	0.80	0.68	0.87	0.91
社会主義○	0.38	0.38	0.55	0.51	0.51	0.66
社会主義○+△	0.54	0.57	0.83	0.67	0.66	0.83
資本主義○	0.33	0.31	0.45	0.46	0.52	0.62
資本主義○+△	0.44	0.38	0.56	0.55	0.59	0.75
合計○	0.33	0.27	0.44	0.38	0.41	0.59
合計○+△	0.51	0.44	0.71	0.59	0.63	0.76