賛成を得やすい文章の機械学習を利用した収集と分析

三木 謙志 [1] 村田 真樹 [2] 馬 青 [3]
[1] 鳥取大学 工学部　電気情報系学科
[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]{b18t2111b@edu.murata@}tottori-u.ac.jp
[3]qma@math.ryukoku.ac.jp

概要

本研究は Yahoo!ニュースのコメント欄を利用して賛成を得やすい文章の特徴の発見を目的とする．賛成した人数とコメント時刻の情報を使用し，同じ記事に対する 2 つのコメントのどちらが賛成を得やすい文章かを機械学習を用いて推定させた．推定させた結果，BERT，ME，SVM の順で正解率が高くなり，一番正解率が高い BERT で 0.7506 となった．素性分析の結果，賛成を得やすい文章の素性だと納得できる素性の発見はできていないが，賛成を得にくい文章の素性だと納得できる素性はいくつか発見した．

1　はじめに

本研究は Yahoo!ニュースのコメント欄を利用して賛成を得やすい文章の特徴の発見を目的とする．賛成した人数とコメント時刻の情報を使用し，同じ記事に対する 2 つのコメントのどちらが賛成を得やすい文章かを機械学習を用いて推定させる．自動分類のみならず，なぜ自動分類できたかの理由を素性分析の技術を利用し，賛成を得やすい文章にどのような特徴があるかを分析する．この分析は賛成を得やすい文章の作成につながると考えている．

2　先行研究

文献 [1] では，どういう文章構造が説得力を持つかを調査している．文献 [2] では，政治のスピーチで説得力のある発話にタグを付与したコーパスを作成している．文献 [3] では，説得を行う文章の構造に関する調査をしている．説得力のある文章は，賛成を得やすい文章と類似するため，これらの研究は本研究に関連する．しかし，これらの研究は，文章構造や文章のパターンに基づき説得力について分析を行っている．一方，端らの研究 [4] では，感動を与える文の作成支援のために感動を与える文の収集とそれらの分析を行った．研究結果では感動を与える文に多く出現する単語として，「人生」，「人々」，「我々」，「恋愛」，「喜び」などが確認された．感動を与えるということは人の感情を揺さぶる単語であるので本研究においても注目すべき単語であると考える. 他にも似たような研究として村田らの研究 [5] [6] [7] も存在している．

3　提案手法

本研究では，最大エントロピー法 (ME)，SVM，BERT の 3 種類の機械学習を利用し，知見獲得を行う．

3.1　ME

最大エントロピー法とは, あらかじめ設定しておいた素性 $f_j(1 \leq j \leq k)$ 集合を F とするとき, 式 (1) を満足しながらエントロピーを意味する式 (2) を最大にするときの確率分布 $p(a,b)$ を求め, その確率分布にしたがって求まる各分類の確率のうち, もっとも大きい確率値を持つ分類を求める分類とする方法である．

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad (1)$$
$$for \quad \forall f_j(1 \leq j \leq k)$$

$$H(p) = -\sum_{a\in A,b\in B} p(a,b)log(p(a,b)) \quad (2)$$

ただし, A, B は分類と文脈の集合を意味し, $g_j(a,b)$ は 文脈 b に素性 f_j があってなおかつ分類が a の場合 1 となり, それ以外で 0 となる関数を意味する. ま

た, $\tilde{p}(a,b)$ は, 既知データでの (a,b) の出現の割合を意味する.

式 (1) は, 確率 p と出力と素性の組の出現を意味する関数 g をかけることで 出力と素性の組の頻度の期待値を求めることになっており, 右辺の既知データにおける期待値と, 左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として, エントロピー最大化 (確率分布の平滑化) を行って, 出力と文脈の確率分布を求めるものとなっている.

3.2 SVM

SVM は，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．このとき，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) が大きいものほどテスト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.

3.3 BERT

BERT は，Bidirectional Encoder Representations from Transformers の略で，「Transformer による双方向のエンコード表現」と訳され，2018 年 10 月に Google の Jacob Devlin らの論文 [8] で発表された自然言語処理モデルである．従来の機械学習では，大量のラベルのついたデータを用意させ，処理を行うことで課題に取り組む．しかし従来の手法に対し，BERT は事前学習でラベルのないデータをはじめに大量に処理を行う．その後，ファインチューニングで少量のラベルのついたデータを使用することで課題に対応させる.

3.4 データ作成

データの作成には Yahoo!ニュースのコメント欄にある，時刻と賛成した人数の情報を利用する．1,000 コメント以上投稿されている記事を対象に，コメントと時刻の情報を収集している．同じ記事に投稿された 2 つのコメントにおいて，コメント時刻がより最近，賛成した人数がより多いという 2 点を満たすコメントを賛成を得やすい文章，もう一方を賛成を得にくい文章だと定義する．そのような賛成を得やすい文章と賛成を得にくい文章を文章対として大量に作成する．ただし，一度でも文章対を作成する際に使用されたコメントは他の文章対では使用しないこととする．例を以下に掲載する.

- コメント A，賛成:4,234, コメント時刻:4 時間前
- コメント B，賛成:3,823, コメント時刻:2 時間前
- コメント C，賛成:6,923, コメント時刻:8 時間前
- コメント D，賛成:2,182, コメント時刻:5 時間前

コメント A を対の一方として文章対を作成する場合，定義に当てはめると文章対として使用できるもう一方はコメント D のみである．コメント B を対の一方として利用する場合も同様にコメント D のみであり，コメント C は高評価数は一番多いがコメント時刻が一番古いためどのコメントとも文章対を作成することはできない.

よってこの 4 つのコメントで定義に当てはまり，文章対として使用できるコメント対はコメント A とコメント D，コメント B とコメント D のみである．しかし，本研究では文章対作成の際に一度使用したコメントは利用しないので文章対として使用できるのはどちらか一つとなっている．このように定義にあてはまる文章対をコメントデータから大量に作成する.

3.5 推定方法

作成した文章対の一方の文章を左側，もう一方を右側として，左側の文章に"L"を右側の文章に"R"を付与したものを学習データとする．文章対を入力とし，"L"，"R"を出力とする．入力が与えられるとそれに対する出力を推定できるように機械学習 (最大エントロピー法 (ME)，SVM，BERT) で学習する.

3.4 節の文章対を学習データとした場合の例を以下に示す．例のように作成した文章対に"L"と"R"と，それに対し文章を反転させたものも同時に学習データとして使用する.

- L, コメント A, コメント D
- R, コメント D, コメント A

例のような学習データを機械学習で右側の文章対の場合に左側の"L" or "R"となるように学習する．そして別の右側の文章対を入力しその場合の"L" or "R"を推定する．左側の文章が賛成を得やすいと判断した場合は"L"を，右側の場合は"R"を出力する.

3.6 素性

ME，SVM では文章対において左側の文章にある単語は「L：単語」，右側にある単語は「R：単語」と「コメントの文字数」を素性として利用する．BERT に入力する際は素性の入力は不要なので文章対のみ

を入力している．

「コメントの文字数」の素性は「L：〜以下」,「L：〜より大きい」,「R：〜以下」,「R：〜より大きい」としおり，〜には「10」,「20」.「50」,「100」,「200」,「500」,「1,000」のいずれかの数字がコメントの文字数に応じて入る．また,「コメントの文字数」を素性とせずに単語だけを素性とする実験も行っている．ME は正規化 α 値，SVM は分離平面を用いて素性分析を行う．ME では，正規化 α 値の高いものが重要な素性となる．SVM では「L：単語」などの 1 単語を入力し，分離平面からの距離が大きいものが重要な素性となる．文章対を用いた BERT の素性分析は困難なため行っていない．

4 実験

4.1 推定実験

1,000 コメント以上を持つ記事を 300 記事分収集し，同記事内の賛成数とコメント時刻の 2 つの条件を満たすコメントから文章対を作成している．対とするコメントの組み合わせは条件を満たすコメントからランダムに決定している．作成した 16,342 組の文章対のうち，8,172 組を学習データ，8,170 組をテストデータとする．BERT では 8,172 組の学習データのうち，2,045 組を検証データ，6,129 組を訓練データとして実験を行っている．

表 1 は機械学習により"L" or "R"のどちらの文章が賛成を得やすいかを推定したときの正解率を示している．

表 1 機械学習の性能評価

文字数素性	有	無
ME	0.7215	0.7209
SVM	0.6788	0.6734
BERT		0.7506

表 1 に示されているように BERT，ME，SVM の順で正解率が高いことがわかる．一番性能が高い BERT で 0.7506 という正解率を得た．また，今回の実験では ME，SVM ともに文字数の素性の有無で性能が大きく変化することはなかったが文字数の素性を利用するほうが正解率がわずかに高くなった．

4.2 ME の素性

ME の素性分析の結果で得られた素性の上位 10 個を表 2 と表 3 に示す．

表 2 ME の素性

賛成を得やすい素性	賛成を得にくい素性
うーん	バカ
予防	ケース
人材	わ
現場	死刑
杏	バラマキ
任意	基準
進ん	こいつ
類	幼稚
持た	理不尽
宮内庁	育成

表 3 ME の素性 (文字数素性無し)

賛成を得やすい素性	賛成を得にくい素性
うーん	バカ
人材	ケース
現場	わ
予防	死刑
任意	幼稚
杏	育成
宮内庁	理不尽
進ん	こいつ
持た	奴
類	やん

表 4 ME の素性 (文字数素性のみ)

賛成を得やすい素性	賛成を得にくい素性
L100 より大きい	L100 以下
R100 より大きい	R100 以下
R20 以下	R20 より大きい
L20 以下	L20 より大きい
L200 より大きい	L200 以下
R200 より大きい	R200 以下
R50 以下	R50 より大きい
L50 以下	L50 より大きい
L10 より大きい	L10 以下
R10 より大きい	R10 以下

表 2 と表 3 から文字数の素性の有無では賛成を得やすい文章の素性にはあまり変化が見られなかった．ME の文字数素性だけを抜き出したのが表 4 である．表 4 からは 100 字より多い文字数の文章が一番賛成を得やすい文量だと読み取れる，

4.3 SVM の素性

SVM の素性分析の結果で得られた素性の上位 10 個を表 5 と表 6 に示す.

表 5 SVM の素性

賛成を得やすい素性	賛成を得にくい素性
予防	幼稚
現場	基準
認める	創価学会
任意	バカ
限界	泥棒
うーん	問う
すみ	期日
子育て	育成
経費	ゴリ押し
製品	失脚

表 6 SVM の素性 (文字数素性無し)

賛成を得やすい素性	賛成を得にくい素性
現場	幼稚
予防	基準
任意	バカ
認める	泥棒
限界	問う
すみ	余っ
うーん	創価学会
子育て	期日
意地	失脚
製品	育成

表 7 SVM の素性 (文字数素性のみ)

賛成を得やすい素性	賛成を得にくい素性
L200 より大きい	L200 以下
R200 より大きい	R200 以下
L100 より大きい	L100 以下
R100 より大きい	R100 以下
L20 以下	L20 より大きい
R20 以下	R20 より大きい
L50 以下	L50 より大きい
R50 以下	R50 より大きい
L10 より大きい	L10 以下
R10 より大きい	R10 以下

表 5 と表 6 から読み取れるように SVM も ME と同様に文字数の素性の有無では賛成を得やすい文章の素性にはそこまで変化が見られなかった．SVM の文字数素性だけを抜き出したのが表 7 である．表 7 からは 200 字より多い文字数の文章が一番賛成を得やすい文量だと読み取れる，

4.4 素性分析

本研究では素性分析を用いて重要素性の獲得をする．4.2 節と 4.3 節の表 2 から表 7 に ME，SVM の機械学習で獲得できた上位 10 素性を掲載している．良い素性とされている「うーん」,「予防」,「現場」,「任意」などは上位 10 素性に共通して現れているが，これらの単語が要因で賛成を得ているとは考えにくい．逆に悪い素性に共通して現れている「バカ」,「幼稚」などは抽象的に誰かを批判する単語なのでこれらの単語が含まれている文章は内容がなく，賛成を得にくいのではないかと考えている．

また，ME では 100 文字より大きい文字数の文章が賛成を得やすいとされ，SVM では 200 文字より大きい文字数の文章が賛成を得やすいとされていることから，賛成を得やすい文章にはある程度の文字数が必要になるのではないかと考えている．

5 おわりに

本研究では賛成を得やすい文章の特徴の発見を目的としている．Yahoo!ニュース内にある 1,000 コメント以上持つ 300 記事からコメント時刻と賛成した人数の情報を使用し，同じ記事に対する 2 つのコメントを比較する．その際にコメント時刻がより最近であるコメント，賛成した人数がより多いコメントの 2 点を満たすコメントを賛成を得やすい文章だと定義する．この 2 点を満たすコメントで文章対を大量に作成し，どちらの文章が賛成を得やすいかを教師あり機械学習を利用し推定させた．推定させた結果，BERT，ME，SVM の順で正解率が高くなっている．一番性能が高い BERT の正解率が 0.7506 となった．素性分析では，賛成を得やすいとされた素性に納得できる単語はなかったが，賛成を得にくいとされた素性には納得できる素性が存在した.

今後は分野によって良い素性は変化するかなどの分野依存性や高評価だけでなく定義の部分で低評価数も考慮した賛成を得やすい文章の分析を行い，今回の研究と比較し，違いの発見などを目指す.

参考文献

[1] 石黒圭. 日本語学習者の作文における文章構成と説得力の関係. 一橋大学国際教育センター紀要, Vol. 8, pp. 3–14, 2017.

[2] Marco Guerini, Carlo Strapparava, and Oliviero Stock. Corps: A corpus of tagged political speeches for persuasive communication processing. **Journal of Information Technology & Politics**, Vol. 5, No. 1, pp. 19–32, 2008.

[3] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 46–56, 2014.

[4] 端大輝, 村田真樹, 徳久雅人. 感動を与える文の自動取得と分析. 言語処理学会第 18 回年次大会, pp. 303–306, 2012.

[5] 村田真樹, 西村涼, 金丸敏幸, 土井晃一, 松岡雅裕, 井佐原均. 情報の重要度を決める要因の抽出・分析と重要度の自動推定. 言語処理学会第 14 回年次大会, pp. 907–910, 2008.

[6] 村田真樹, 西村涼, 金丸敏幸, 土井晃一, 鳥澤健太郎. ユーザ個人の興味の影響を考慮した情報の重要度を決める要因の抽出・分析. 言語処理学会第 15 回年次大会, pp. 554–557, 2009.

[7] 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, Volume 2, Issue 4**, pp. 272–279, 2010.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.