Web 上のレビュー記事のイデオロギー分析とその応用

橋本悠　掛谷英紀
筑波大学

概要　国会議事録を教師信号とした学習を行うことで、文章を政治的イデオロギー別に分類する研究が行われている。しかし、同研究ではあるイデオロギーを否定している文章がそのイデオロギーとして判定されるという問題が発生している。そこで本研究では、大手ポータルサイト Yahoo！JAPAN が運営する政治評価サイト「Yahoo!みんなの政治」内のコンテンツ「みんなの評価」と呼ばれるユーザー投稿型の国会議員レビューを教師信号とすることによって、肯定意見・否定意見を考慮に入れたイデオロギー分類システムを構築する。また、同システムに Amazon.com の新書のブックレビューを入力し、イデオロギーに則した図書の推薦を試みる。

1　はじめに

これまで、文章をイデオロギー別に分類する研究が畑中らによって行われている。畑中らは、イデオロギーを測る客観的指標として、これまで大手新聞社や政党に注目してきた[1,2]。具体的には、新聞社説や国会議事録を機械学習にかけ、それに基づき文章のイデオロギー別分類を行っている。しかしこの手法はその性質上、否定意見と肯定意見の考慮をしておらず、あるイデオロギーに言及しながらそれを否定している文章が、そのイデオロギーに属する文章であると判定される傾向があり、それが課題となっている。

そこで本研究では、肯定意見、否定意見という違いも考慮に入れたイデオロギーを測る客観的指標として Yahoo!JAPAN が運営する政治評価サイト「みんなの政治」内の国会議員のユーザー評価記事に着目する[3][4]。その中から「民主党」と「自民党」の議員へのユーザー評価記事を教師信号とし、政党の左右から文章をイデオロギー別に分類するシステムの構築を試みる。またシステムの分類の正当性を、学習の結果得られたパラメータおよびクロスバリデーションの正答率による評価、文章の書き手による判定結果の主観評価、という 2 つの側面から検証する。また応用として、学習したシステムを用い Amazon.com の新書ブックレビューをテストすることで、新書の分類結果から、新書と政党の関連性の検証を行う [5]。

2　システムの概要

本研究では、形態素解析ツールとして、ChaSen を用いる[6]。まず ChaSen を用いて文書データを形態素解析し、品詞ごとに単語を分割して素性として抽出し、それらから学習データ及びテストデータを作成する。

学習データを元に、機械学習のプログラムで文章の特徴を学習し、テストデータを元にシステムの精度を算出する。機械学習には最大エントロピー法を用いる[7]。システムの概要を図 1 に示す。最大エントロピー法のプログラムとしては maxent を利用する[8]。

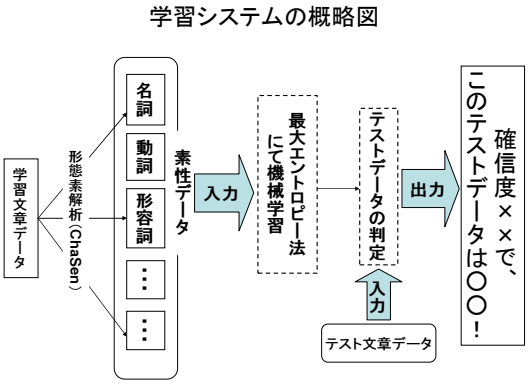

図 1　システムの概要

学習に使用する単語は限定せず、学習データに現れる全ての単語を原則用いることにする。畑中の研究によってイデオロギーの学習では、素性は名詞と動詞、および形容詞と名詞で組み合わされた熟語を用いれば良好な結果が得られることが示されている。しかし今研究ではイデオロギーと共に肯定・否定の違いも学習させなければいけなく、肯定・否定の違いがどのような単語として現れるのか現時点では判断できないため、全単語を用いて学習を行う。

3　ユーザー議員評価の学習

3.1　「みんなの評価」の分析

本研究の学習では国内大手ポータルサイト Yahoo!JAPAN が運営する政治評価サイト「みんなの政治」における、一般ユーザーが任意の議員の評価記事を書くことのできるサービス「みんなの評価」を用いる。これに投稿されているユーザー議員評価から、民主党か自民党に所属している国会議員のユーザー議員評価記事を抜き出して利用する。「みんなの政治」に登録されている国会議員は現行の国会議員のみであり、本研究では 2009 年 9 月時点での現行の民主党、自民党に所属している国会議員の評価記事を入手した。また、現行の国会議員であれば、最長で 2007 年 6 月 6 日以降に投稿された評価記事ならば参照することができるので、集めた評価記事は 2007 年 6 月 6 日から 2009 年 9 月までのものとなる。図 2 に実際の議員評価記事の一例を示す。

図２　議員評価記事の一例

図 2 はある議員に対してユーザーによって書かれた議員評価記事である。評価記事は全て図 2 の形式に沿って書かれている。議員評価記事は評価点・タイトル・本文と詳細グラフによって構成されている。評価点は星 1～星 5 までつけることができ、星 1 が最低評価、星 5 が最高評価となる。ここで、評価点が低い記事の内容はその議員に対しての否定的な意見としてみなすことができ、逆に評価点の高い記事の内容は肯定的な意見とみなすことができる。

そこで、本研究では肯定意見、否定意見という違いも考慮に入れたイデオロギーを分析することを目的としているので、集めた文章から評価点の低い記事と評価点の高い記事だけを抽出し、それらを学習データに採用する。

また、議員評価記事を抽出する際に、評価記事中の文章が短すぎることを考慮して、学習に使用する記事の制約として記事中の文字数が電子データとして 600 バイト以上のものであるという条件を加味する。評価記事を投稿する際にサイト上では評価記事に最低文字制限を設けていないため、評価記事が 1 文や一言のものが多い。このようなあまりにも短い評価記事は意見として成立しているとはいえないので、このような評価記事は学習データとしてふさわしくない。

この条件で抽出した自民党・民主党各党の高評価記事・低評価記事の集計結果が図 3 である。

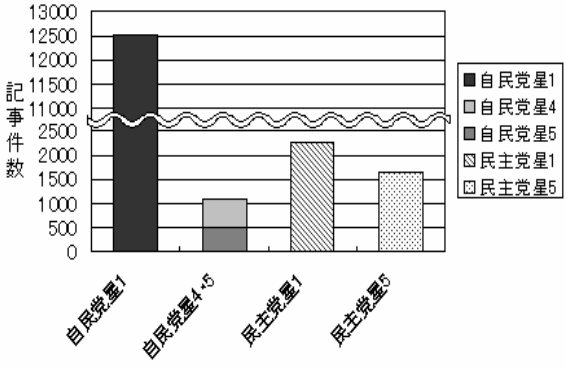

図３　議員評価記事の抽出件数

ここで、「自民党星 1」とは自民党の議員に対して星 1 の評価点をつけている評価記事の件数を表している。他のラベルも同様である。図 3 を見てもらえればわかるように、抽出記事件数に大きな偏りがあることがわかる。特に自民党の議員に対する星 1 の評価記事件数は圧倒的である。また自民党の星 5 の評価記事も極端に少ない。学習する際に、カテゴリによって発言の件数に差が大きいと、判定結果が件数の多いカテゴリに近づいてしまう。そこで、一番件数の少ない自民党星 5 にあわせることを考えたが、それでは学習データ数が少なすぎる。よって星 4 の評価記事も加えた、自民党星 4･5 の評価記事件数にあわせることにする。ここからさらに、自民党星 4･5 の評価記事 1099 件を、1 人の議員に対する評価記事が全記事件数の 20%以下となるよう整理すると、評価記事件数は 950 件となった。よって他のカテゴリからも 950

件の発言を同様の条件でランダムに選び、これらを学習データに採用し、学習・実験を行う。

3.2 クロスバリデーションによる判定精度の検証

「自民党星 1」「自民党星 4･5」「民主党星 1」「民主党星 5」それぞれ 950 件の発言を学習し、10 分割のクロスバリデーションにて判定精度の検証を行ったところ、正解率は 65.08%となった。なお、党名・人名はイデオロギーを反映している素性とはいえないため排除して学習を行っている。

各カテゴリ別に、どのカテゴリの発言と分類されたかの割合を図４に示す。図の上から順に、自民党星 1 の評価記事をテストした場合、自民党星 4･5 の評価記事をテストした場合…となっており、グラフの部分が学習結果によってどのカテゴリの発言と判断されたかの件数の割合を示している。

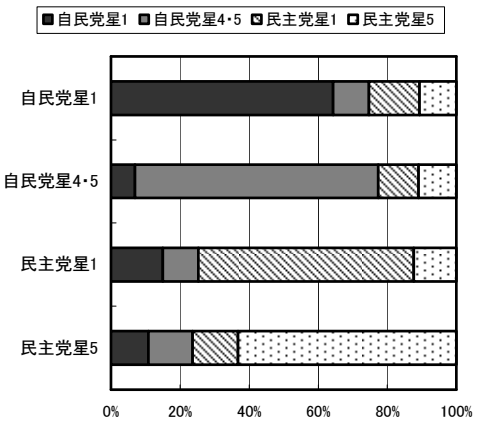

図４　学習結果　カテゴリ別正解率

全体的に見ると、どのカテゴリも同程度の正解率となっている。学習が全く上手く行かなかった場合に正解率が 25%付近となることを考慮すると、今回の学習はある程度各カテゴリの特徴であるイデオロギーの差異を把握できているといえる。

ここで、判定システムがどういった素性を手がかりにして政党別に発言を分類しているかを見てみるため、政党別に判定に大きな影響を与えた上位の素性をいくつかピックアップする。（表１）

表１　判定に影響を与えた素性の一部抜粋

自民党星1	自民党星4･5	民主党星1	民主党星5
補正	死刑	外国	若手
世襲	北朝鮮	在日	圧勝
アメリカ	靖国	韓国	交代
スポーツ	中国	日教組	薬害
自衛隊	外交	マルチ	活躍
高齢	左翼	反日	少子化
医療	マスコミ	自治労	信頼
反省	頑張っ	パチンコ	真面目
疑惑	評価	不安	すばらしい
即刻	成果	疑問	優秀
不正	正しい	マイナス	自公

3.3 主観評価による判定精度の検証

判定システムの示す結果がどの程度妥当であるかを検証するために、政治話題に明るい大学教員Ｋ氏に協力を得て主観評価を行う。Ｋ氏が過去に書き溜めた時事問題や社会問題に関するエッセイをテストデータとして計 108 件入力し、それぞれのエッセイについて、出された判定結果がどの程度妥当であるかをＫ氏本人に４段階で評価させる。その結果、「正当 75 件」「許容 23 件」「やや不当 10 件」「不当 0 件」との主観評価が得られた。「正当」、「許容」と評価されたものが全体の 9 割程度となっている。畑中らの研究においても同様のエッセイでの主観評価をＫ氏に行わせているが、そのときは「正当」、「許容」と評価されたものが全体の７割程度であった[1]。このことを考慮すると、クロスバリデーションの結果はあまり振るわなかったものの、実際の判定では高い精度での判定が行えていると考えられる。今後様々な文章のテストを行い、それらでも同様な結果が得られるのか検証していく必要がある。

4　ブックレビューの検証

3 章により 4 カテゴリの議員評価記事を学習したシステムに、Amazon.com のブックレビューをテストにかける実験を行う。

Amazon.com のブックレビューは「みんなの評価」と形式がほぼ同様に、評価点、タイトル、本文によって構成されており、今回はその中から新書のブックレビューを用いて検証を行う。

使用する新書のブックレビューは、岩波新書・講談社現代新書・集英社新書・新潮新書・中公新書・文春新書から発刊された本のうち、2009 年 7

月までに掲載されていたものを用いている。そこからさらに肯定的意見である評価が高い星 4･5 のブックレビューと、否定的意見である評価が低い星 1･2 のブックレビューを抽出し、テストを行う。

テストでは各ブックレビューに対して、「そのレビューが各カテゴリである確率」が出力される(その中で最高確率であるカテゴリが判定結果となる)。

そこで、そこから新書ごとに星 4･5 のブックレビューに対する出力から自民党星 4･5、民主党星 5 の各カテゴリの確率の幾何平均を求めることで、肯定的なブックレビューが自民党星 4･5 と判定される確率が高い新書と、反対に民主党星 5 と判定される確率が高い新書を調べた。同様に星 1･2 のブックレビューが自民党星 1 と判定される確率が高い新書、民主党星 1 と判定される確率が高い新書も調べた。そこから上位に挙がった特徴的な新書をリストアップしたものが表 2 である。

表 2　各カテゴリ上位に挙がった新書リスト

星4･5のブックレビューに関して

自民党星4･5	民主党星5
･司法は腐り人権滅ぶ ･昭和天皇の履歴書 ･憲法と国家一同時代を問う ･日米同盟の正体~迷走する安全保障 ･「権力社会」中国と「文化社会」日本 ･北朝鮮の外交戦略	･私は女性にしか期待しない ･男女交際進化論 ･住まいと家族をめぐる物語一男の家、女の家、性別のない部屋 ･男と女の法律戦略 ･不思議の国アメリカー別世界としての50州 ･中国革命を駆け抜けたアウトローたち

星1･2のブックレビューに関して

自民党星1	民主党星1
･夫婦の格式 ･源氏と日本国王 ･民主党一野望と野合のメカニズム ･米軍再編と在日米軍 ･金正日の正体 ･貧困の克服一アジア発展の鍵は何か ･堂々たる政治	･憲法「押しつけ」論の幻 ･南京事件一「虐殺」の構造 ･「在日」としてのコリアン ･ローマ教皇とナチス ･親米と反米一戦後日本の政治的無意識 ･マッカーサー元帥と昭和天皇

自民党と民主党のイデオロギーの違いを思想の左右の視点から考えると、自民党は右寄りのイデオロギーであり、民主党は左寄りのイデオロギーとみなすことができる[9]。

そのことを考慮して表 2 を見てみると、肯定的(星 4･5)ブックレビューが自民党星 4･5 の議員評価に近いと判定された新書はそのタイトルからも右寄りの思想を持つ書籍が多いことがわかる。また、民主党星 5 の議員評価に近いと判定された新書には、左寄りの思想であると予想される書籍や、フェミニスト的なタイトルである書籍が多い。

否定的(星 1･2)ブックレビューでは、自民党星 1＝左寄りだと判定されているものには左寄りの思想を持つ人が嫌う（低評価する）右寄りの思想の新書が多く、民主党星 1＝右寄りだと判定されているものには、右寄りの人が嫌うと思われる左寄りの内容と想起されるタイトルの新書が挙がっており、イデオロギーに則った書籍が選別されている。

これらのことから、学習システムによって得られた、自民党・民主党のイデオロギーの差異にのっとった書籍が提示されていることがわかる。

5　おわりに

本研究ではイデオロギーを測る客観的指標として政党の左右に着目し、新たに肯定意見・否定意見という概念を考慮に入れた教師信号を用いることで文章をイデオロギー別に分類すること試みた。本稿では政治評価サイト「みんなの政治」内の「みんなの議員評価」と呼ばれる議院評価記事を教師信号として学習し、学習結果の有効性を確認した。

また、学習結果を用い Amazon.com の新書ブックレビューをテストすることによって新書と政党の関連性を検証し、システムの分類精度を評価したところ、この手法によって言論のイデオロギー別分類が可能であることを示唆する良好な結果が得られた。

今後の予定としては、ブックレビュー以外の様々な言論と政党の関連性の検証を行うことを考えている。また、判定精度の向上も今後の課題である。

参考文献

[1] 畑中允宏, 村田真樹, 掛谷英紀 (2009): 新聞社説・国会議事録に基づく言論のイデオロギー別分類, 言語処理学会第 15 回年次大会発表論文集

[2] 畑中允宏, 金丸敏幸, 村田真樹, 掛谷英紀 (2008):新聞の社説を教師信号とする文章の右翼度・左翼度判定, 言語処理学会第 14 回年次大会発表論文集

[3] Yahoo!JAPAN http　http://www.yahoo.co.jp/

[4] Yahoo!みんなの政治　http://seiji.yahoo.co.jp/

[5] Amazon.com　http://www.amazon.co.jp/

[6] 奈良先端科学技術大学院大学 松本研究室 ChaSen http://cl.aist-nara.ac.jp/

[7] Ristad, E. S. (1998). "Maximum Entropy Modeling Toolkit, Release 1.6 beta."1997 http://www.mnemonic.com/

[8] 内山将夫氏. maxent　http://www2.nict.go.jp/x/x161/members/mutiyama/software.html

[9] Wikipedia 英語版 - List of political parties in Japan